Understanding human dynamics from large-scale location-centric social media data : analysis and applications

Dingqi Yang

▶ **To cite this version:**

Dingqi Yang. Understanding human dynamics from large-scale location-centric social media data : analysis and applications. Other [cs.OH]. Institut National des Télécommunications, 2015. English. NNT : 2015TELE0002 . tel-01115101v3

HAL Id: tel-01115101

https://theses.hal.science/tel-01115101v3

Submitted on 24 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE DE DOCTORAT DE TELECOM SUDPARIS et L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité : Informatique

Ecole doctorale : Informatique, Télécommunications et Electronique de Paris

Présentée par

Dingqi YANG

Pour obtenir le grade de
DOCTEUR DE TELECOM SUDPARIS

Exploration de la Dynamique Humaine Basée sur les Données Massives des Réseaux Sociaux de Géolocalisation : Analyse et Applications

Soutenue le 27 Janvier 2015

Devant le jury composé de :

Eric Gaussier	Rapporteur	Professeur, Université Joseph Fourier - Grenoble, France
Daniel Gatica-Perez	Rapporteur	Professeur, EPFL - Lausanne, Suisse
Pierre Sens	Examinateur	Professeur, UPMC – Paris, France
Marie-Aude Aufaure	Examinateur	Professeur, Ecole Centrale Paris - Paris, France
Cécile Bothorel	Examinateur	Maître de conférence, Institut Mines-Télécom – Brest, France
Djamal Zeghlache	Directeur de thèse	Professeur, Institut Mines-Télécom – Evry, France
Daqing Zhang	Co-encadrant	Directeur d'études, Institut Mines-Télécom – Evry, France

Thèse N° : 2015TELE0002

Doctor of Philosophy (PhD) Thesis
Université Pierre & Marie Curie -TELECOM SudParis

Specialization

Informatique

Presented by

Dingqi YANG

Submitted for the partial requirement of

Doctor of Philosophy
from
Université Pierre & Marie Curie (UPMC) - TELECOM SudParis

Understanding Human Dynamics from Large-Scale Location-Centric Social Media Data: Analysis and Applications

January 27, 2015

Committee:

Éric Gaussier	Reviewer	Professor, Université Joseph Fourier - Grenoble, France
Daniel Gatica-Perez	Reviewer	Professor, EPFL - Lausanne, Switzerland
Pierre Sens	Examiner	Professor, UPMC - Paris, France
Marie-Aude Aufaure	Examiner	Professor, Ecole Centrale Paris - Paris, France
Cécile Bothorel	Examiner	Associate Professor, Institut Mines-Télécom - Brest, France
Djamal Zeghlache	Thesis Director	Professor, Institut Mines-Télécom/Télécom SudParis - Evry, France
Daqing Zhang	Advisor	Professor, Institut Mines-Télécom/Télécom SudParis - Evry, France

Declaration

This thesis:

— is the result of my own research work and contains nothing which is the outcome of work done in collaboration with others, except where specifically indicated in the text;

— has not previously been submitted for a degree or diploma, or other qualification at any other university.

Dingqi Yang
January 2015

Abstract

Human dynamics is an essential aspect of human-centric computing which is a transdisciplinary research field combining human factors and computer science. Studying human dynamics focuses on understanding the underlying patterns, relationships, and changes of human behavior. By analyzing human dynamics, we can understand not only individual's behavior, such as a presence at a specific place, but also collective behavior, such as crowd mobility and social movement. Understanding human dynamics can thus enable various applications, such as personalized location based services in smart city scenarios. However, before the availability of the ubiquitous smart devices (e.g., sensor-embedded smartphones), it is practically difficult to collect large-scale human behavior data.

With the ubiquity of GPS-equipped smartphones, location-centric social media, i.e., location based social networks (LBSNs), has gained increasing popularity in recent years, which makes large-scale user activity data become attainable. In LBSNs, users can share their real time activities with their friends by checking in at points of interests (POIs), such as a restaurant or a bar. Such location-centric social media data massively implies human dynamics. For example, from individual perspective, we can explore spatial temporal regularity of user activities; from collective perspective, we can investigate the collective behavior patterns and study their difference across societies.

In this dissertation, we explore human dynamics based on big location-centric social media data, and investigate into the whole life-circle of the research process, including data collection, analysis and applications. Concretely, in order to collect large-scale user activity data, we first build a platform to collect user activity data from various LBSNs, such as Foursquare and Twitter. Based on this location-centric social media data, we then study human dynamics and their applications from both individual and collective perspectives.

From individual perspective, based on city-scale user activity data, we explore user preference on POIs and the spatial-temporal regularity of user activities. Specifically,

- In order to study user preference with different granularity and its applications in service personalization, we define and extract two types of user preference, viz., coarse-grained user preference (i.e., user-POI preference) and fine-grained user preference (i.e., user-POI-item preference), from heterogeneous user activity data in LBSNs (e.g., check-ins and user's comments). To incorporate these two types of user preference into personalized location based services, we propose a preference-aware POI recommendation and search framework by designing two novel algorithms based on low-rank approximation techniques for efficient user preference prediction.
- In order to study the spatial-temporal regularity of user activities and its applications in activity preference inference tasks, we propose a novel spatial temporal activity model, which can efficiently capture spatial and temporal patterns of user activity from the sparse check-in data. For spatial patterns, we propose the notion of personal functional region and related parameters to model and infer user spatial activity preference. For temporal patterns, we exploit the temporal activity similarity among

different users and apply non-negative tensor factorization to collaboratively infer temporal activity preference. Finally, we put forward a context-aware fusion framework to combine the spatial and temporal models for accurate activity preference inference.

From collective perspective, based on global-scale user activity data, we study the collective activity pattern with both country and city granularity, and its correlation with global cultures.

- In order to study the nation-wide collective behavior, we develop NationTelescope, a platform that monitors, compares, and visualize large-scale collective behavior in LBSNs. It is designed to let user efficiently explore the behavioral differences across countries. To achieve this goal, we leverage a slide-window based approach to detect the discriminative activities according to the related traffic patterns in different countries, and implement an interactive map interface for data visualization.
- In order to study the correlation between collective behavior and human cultures on a global scale, we investigate into the city-wide collective behavior, and propose a participatory cultural mapping approach to automatically discover the cultural clusters of cities and generate a cultural map. Specifically, since only local users are eligible for representing local cultures, we propose a progressive "home" location identification method to filter out ineligible users. By extracting three key cultural features from daily activity, mobility and linguistic perspectives respectively, we propose a cultural clustering method based on spectral clustering techniques to discover the cultural clusters of cities.

Finally, we summarize our findings with regard to individual and community dynamics and discuss potential future research trends, such as privacy issues of such location-centric social media data, combination of human activity data and various ubiquitous sensor data, the big data challenges of processing such large-scale human activity data and more innovative applications in smart city scenarios.

Keywords

Human dynamics, Social media analysis, Location based social networks, Location based services, Participatory sensing, Recommendation system, Personalized search, Sentiment analysis, Spatial, Temporal, Collective behavior, Cultural analysis.

Résumé

La dynamique humaine est un sujet essentiel de l'informatique centrée sur l'homme, domaine de recherche transdisciplinaire combinant les facteurs humains et l'informatique. L'étude de la dynamique humaine se concentre sur la compréhension des régularités sous-jacentes, des relations, et des changements dans les comportements humains. En analysant la dynamique humaine, nous pouvons comprendre non seulement des comportements individuels, tels que la présence d'une personne à un endroit précis, mais aussi des comportements collectifs, comme la mobilité de la foule et les mouvements sociaux. L'exploration de la dynamique humaine permet ainsi diverses applications, entre autres celles des services géo-dépendants personnalisés dans des scénarios de ville intelligente. Cependant, avant la disponibilité des appareils intelligents omniprésents (p. ex., les smartphones avec capteurs embarqués), il était pratiquement impossible de recueillir des données sur le comportement humain à grande échelle.

Avec l'omniprésence des smartphones équipés de GPS, les réseaux sociaux de géolocalisation ont acquis une popularité croissante au cours des dernières années, ce qui rend les données d'activité des utilisateurs disponibles à grande échelle. Sur les dits réseaux sociaux de géolocalisation, les utilisateurs peuvent partager leurs activités en temps réel avec leurs amis par l'enregistrement de leur présence (c.-à-d., des check-ins) à des points d'intérêt (POIs), tels qu'un restaurant ou un bar. Ces données d'activité enregistrées par les utilisateurs contiennent des informations massives sur la dynamique humaine. Par exemple, du point de vue individuel, nous pouvons explorer la régularité spatio-temporelle des activités des utilisateurs ; et du point de vue collectif, nous pouvons étudier les comportements collectifs et comprendre certaines différences culturelles entre les diverses sociétés.

Dans cette thèse, nous explorons la dynamique humaine basée sur les données massives des réseaux sociaux de géolocalisation, et étudions toutes les étapes du processus de recherche, y compris la collection et l'analyse de données, ainsi que les applications qui en peuvent en découler. Concrètement, afin de recueillir des données d'activité d'utilisateurs à grande échelle, nous construisons une plate-forme de collecte de données sur différents réseaux sociaux de géolocalisation, p. ex., Foursquare et Twitter. En nous basant sur ces données, nous étudions ensuite la dynamique humaine et ses applications selon des perspectives individuelles et collectives.

Du point de vue individuel, en utilisant les données d'activités d'utilisateurs à l'échelle de la ville, nous explorons les préférences de l'utilisateur quant aux POIs et la régularité spatio-temporelle des activités des utilisateurs. Plus spécifiquement :

- Afin d'explorer la préférence de l'utilisateur avec des granularités différentes et ses applications, en analysant des données hétérogènes d'activité d'utilisateurs (p. ex., les check-ins et les commentaires des utilisateurs), nous définissons deux types de préférences de l'utilisateur : celle avec une granularité grossière (c.-à-d., la préférence utilisateur-POI) et celle avec une granularité fine (c.-à-d., la préférence utilisateur-POI-entité). Pour intégrer ces deux types de préférences de l'utilisateur dans les

services géo-dépendants personnalisés, nous proposons un framework de recommandations et de recherches personnalisées de POI, y compris deux nouveaux algorithmes basés sur des techniques d'approximation de rang réduit pour prédire les préférences de l'utilisateur.

- Afin d'explorer la régularité spatio-temporelle des activités des utilisateurs et ses applications, nous proposons un nouveau modèle spatio-temporel des activités des utilisateurs, capable de modéliser efficacement les caractéristiques spatiale et temporelle des activités des utilisateurs à partir de données clairsemées. Pour la caractéristique spatiale, nous introduisons la notion de région fonctionnelle personnelle afin de modéliser la préférence d'activité spatiale des utilisateurs. Pour la caractéristique temporelle, nous exploitons la similitude temporelle d'activité entre les utilisateurs et appliquons les techniques de décomposition tensorielle non-négative afin de modéliser la préférence d'activité temporelle des utilisateurs. Enfin, nous proposons une méthode combinant les modèles spatiaux et temporels pour inférer précisément la préférence des activités des utilisateurs.

Du point de vue collectif, en utilisant les données d'activités d'utilisateurs à l'échelle globale, nous explorons la forme d'activité collective avec les granularités de pays et ville, ainsi qu'en corrélation avec les cultures. Plus précisément :

- Afin d'analyser les comportements collectifs dans tel pays, nous proposons Nation-Telescope, une plate-forme qui surveille, compare, et visualise les comportements collectifs à grande échelle. Il est spécialement conçu pour permettre à l'utilisateur d'explorer efficacement les différences comportementales collectives entre les pays. Pour atteindre cet objectif, nous proposons une approche basée sur une fenêtre glissante afin de détecter les activités discriminatives entre différents pays. De plus, nous mettons également en œuvre une interface graphique interactive pour la visualisation.
- Afin de comprendre la corrélation entre les comportements collectifs et les cultures humaines à l'échelle globale, en analysant les comportements collectifs de la ville, nous proposons une approche de cartographie culturelle qui permet de découvrir automatiquement les régions culturelles avec une granularité de ville. Concrètement, puisque seuls les utilisateurs locaux sont éligibles pour représenter les cultures locales, nous proposons une méthode progressive pour identifier l'origine de l'utilisateur, afin d'éliminer les utilisateurs inéligibles. Ensuite, en extrayant trois caractéristiques culturelles principales en relation avec la régularité des activités quotidiennes des utilisateurs, la mobilité inter-cités et la linguistique, respectivement, nous proposons une méthode de regroupement culturel basée sur les techniques de regroupement spectral pour découvrir les régions culturelles.

Enfin, nous résumons nos conclusions concernant les dynamiques individuelles et collectives. Nous examinons également les travaux futurs potentiels, tels que la protection de la confidentialité de ces données sur les réseaux sociaux de géolocalisation, l'intégration des données d'activité humaine avec celles des capteurs ubiquitaires, les nouveaux enjeux du traitement de ces données massives, ainsi que les applications innovantes dans des scénarios de ville intelligente.

Mots-clés

Dynamique humaine, Analyse des réseaux sociaux, Réseaux sociaux de géolocalisation, Services géo-dépendants, Détection participative, Système de recommandation, Recherche personnalisée, Analyse des sentiments, Spatial, Temporel, Comportement collectif, Analyse culturelle.

To my dearest wife Bingqing.

Acknowledgements

First of all, I would like to express gratitude to my supervisors, Prof. Daqing Zhang and Prof. Djamal Zeghlache, for their continuous support. They have been nurturing and advising me throughout my study and their support and guidance have been fundamental to shape my research and focus my efforts. I also want to thank all jury members, especially two reviewers, Prof. Eric Gaussier and Prof. Daniel Gatica-Perez.

I am grateful to those with whom I have co-authored papers over the last three years : Prof. Bin Guo, Prof. Zhiwen Yu, Dr. Zhu Wang, Dr. Zhiyong Yu, Dr. Vincent W. Zheng, Dr. Korbinian Frank, Dr. Patrick Robertson, Edel Jennings, Mark Roddy, Dr. Michael Lichtenstern for their guidance during the progress of my research. In addition, I would like to thank my colleagues with whom I have collaborated in various research projects, particularly Haoyi Xiong, Luca Lamorte, Dr. Daqiang Zhang, Prof. Hervé Debar, Prof. Joaquin Garcia-Alfaro, Dr. Gregory Blanc and Dr. Nizar Kheir for many interesting discussions and great insights.

I would also like to show my gratitude to all my other colleagues and friends for their kindness and help during my stay in Institut Mines-Télécom/Télécom SudParis, particularly Dr. Xiaoping Che, Dr. Bin Li, Dr. Chao Chen, Dr. Lin Sun, Leye Wang, Longbiao Chen, Xiao Han, Dr. Songbo Song, Xin Ma, Hui Wang, Hao Feng. It was a great experience working with these smart people.

Finally, I do not even know how to thank my beloved Bingqing. She has been through this Ph.D. with me, day after day, reading drafts of my papers and listening to unpolished versions of my talks. I hope we will remember these years forever, together. In addition, I thank my parents Zengxuan Yang and Yali Fei for their unwavering belief in me and support of my endeavors. I know that I would not accomplish this journey without their infinite love and support.

My heartfelt thanks to you all.

Dingqi @ Paris, France
January, 2015

Table of contents

Chapter 1

Introduction

Contents

1.1 Background

Understanding human dynamics has traditionally played an important role in various human sciences, including sociology, psychology and physics, etc. As a transdisciplinary research field, its main goal is to study and understand human behavior and explore the potential applications of such knowledge. For example, by studying collective behavior and social movement, we can discover some fundamental issues and potential problems confronting our societies [134].

In the early stage of studying human dynamics, research efforts are mainly focused on understanding the ways in which people behave, learn and communicate from psychological perspective, and studying their diversity across different populations. A notable work published in 1997 by Seagal et al. [126] defines human dynamics as a body of work that identifies and illuminates innate distinctions in the way people function as whole systems that include mental, emotional and physical dimensions. They further develop the concept of "personality dynamics" which summarizes the interaction between these three dimensions, and study its application on strengthening organizational performance (e.g., enhancing creativity, optimizing business relationships) in big companies.

Several years later, with the rapid advancement of information technology, research

on human dynamics gained momentum with novel data sources such as the Internet and wireless sensors. For example, in 2005, A.-L. Barabási published his work [12] on studying human dynamics in Nature. Specifically, based on an email traffic dataset that captures the sender, recipient, time and size of each e-mail, he uses a statistical physics model to explain the long tailed distribution of inter event times which naturally occur in human activity. In addition, in 2004, Pentland et al. from the human dynamics group at the MIT Media Laboratory pioneer the idea of wearable computing to study the patterns of face-to-face interactions within the workplace, and their applications on improving the functioning of the organization [109].

While these works provide some valuable insights on understanding human dynamics, they are often limited to some extent by the used behavior data. Specifically, they only focus on specific types of human behavior (e.g., online communication behavior [12], working related activities [109, 126] or social movement [134]), and usually suffer from the lack of large-scale human daily activity data. However, in practice, it is difficult to monitor and collect large-scale human activity data.

In recent years, with the increasing popularity of ubiquitous smart devices, such as wearable computing devices (e.g., Google Glass [1] and Apple Watch [2]), sensor-embedded smartphones and tablets, users leave a considerable amount of digital footprints in their daily lives, which massively reflect their physical activities. According to data collection schemes, there are two main types of digital footprints, namely, *passive* and *active* digital footprints [88]. On the one hand, *passive digital footprints* are generated by users without deliberate intervention from them. For example, by carrying smartphones in their daily activities, users passively leave massive accelerometer readings. By exploring these digital traces, researchers can identify user activities such as walking, jogging, climbing, etc. [72]. On the other hand, *active digital footprints* are voluntarily generated by users through deliberate posting or sharing of their information. For example, using various social media applications, such as Twitter [3] and Foursquare [4], users actively report their activities by sharing their status with their friends, such as having dinner, shopping or working. Compared to the passive digital footprints which mainly consist of low-level sensor data such as accelerometer readings, the active digital footprints on social media contain rich semantic information about user activity. For example, a user may share a status expressing that she is having a dinner with a friend at a French restaurant in central New York on a Friday night. The active digital footprints on social media have the potential to better study human dynamics.

1. `https://www.google.com/glass/start/`
2. `http://www.apple.com/watch/`
3. `https://twitter.com/`
4. `https://foursquare.com/`

Figure 1.1 – An example of check-in activity in LBSNs.

The recent booming of GPS-embedded smartphones makes a novel type of social media become popular, namely, location-centric social media, which is also known as location based social networks (LBSNs). Specifically, according to the smartphone market share data from the International Data Corporation, in the second quarter of 2014, there were 301.3 million smartphones shipped globally, establishing a new single quarter record [5]. The ubiquity of smartphones makes many social media services accessible from mobile applications. The embedded GPS sensors provide users with the opportunity to share their real-time presences at specific locations with their friends on social media, which creates the idea of location-centric social media. Through location-centric social media, users cannot only interact with their friends by sending messages, sharing photos and posting status, etc., but also interact with physical places (i.e., points of interest or venues [6]) by sharing their real-time presences (i.e., performing check-ins), commenting on places (i.e., leaving tips), etc. Figure 1.1 demonstrates a typical check-in activity in LBSNs, which contains rich information about a user's activity, such as time, geo-location, POI category and a short text expressing the current status of the user. With the increasing popularity of such location-centric social media, large-scale user activity data become attainable. For example, Foursquare, a typical LBSN launched since March 2009, has attracted more than 45 million users globally and contained more than 5 billion check-ins by January 2014, with millions more every day [7]. Such large-scale location-centric social media data provides us with an unprecedented opportunity to study human dynamics.

Studying human dynamics consists of understanding both individual and collective human activities and exploring the potential applications.

♢ From individual perspective, based on user activity data in LBSNs, the goal of studying human dynamics is to reveal individual's activity patterns, understand the impli-

5. `http://www.idc.com/prodserv/smartphone-os-market-share.jsp`
6. a "venue" refers to a point of interest in Foursquare, we do not differentiate these two terms throughout this thesis.
7. `https://foursquare.com/about`

cation of the activity patterns, and explore the potential applications of such knowledge. For example, by studying users' check-in traces in LBSNs, we are able to understand individual's daily life patterns and answer the questions like, what kind of places a user would like to go, what does the user like in specific places, etc. Such knowledge is essential to enable personalized location based services.

♢ From collective perspective, based on the large-scale user activity data in LBSNs, the goal of studying human dynamics is to discover the characteristics and regularities of collective activities, study their differences across different populations, and understand its correlation with other cultural and societal factors. For example, by studying the "food related" collective activities in LBSNs, we are able to discover culinary differences across populations, and its correlation with culture, immigration, religion, etc.

1.2 Thesis Contributions and Chapter Outline

In this dissertation, we explore human dynamics based on large-scale user activity data collected from location-centric social media. Since user activities in LBSNs has been widely studied in recent years, in **Chapter 2**, we first present the existing work in this area and point out the related work of human dynamics, and then identify the potential research challenges. The rest of the dissertation presents our contributions, which involve the whole life-circle of the research process, including data collection, analysis and applications.

♢ In **Chapter 3**, we present a data collection platform which is able to continuously collect large-scale user activity data from different location-centric social media (e.g., Foursquare and Twitter). Specifically, due to the privacy protection policy in social media service providers, user activity streams may not be accessed publicly, such as that in Foursquare and Facebook [8]. Fortunately, Twitter public streams [9], are publicly accessible and contains large-scale user activity data. Moreover, this data usually includes user activity data from other LBSNs (e.g., Foursquare) by providing shortened URLs linked to the original LBSNs. By resolving the URLs, we can obtain the user activity data from the original LBSNs. Therefore, we design and develop a data collection platform that first monitors Twitter public streams, and then identify and extract user activity data from different LBSNs. It can automatically collect user activity data from different LBSNs in a streaming manner. In addition, by investigating into the specific characteristics of user activity data in LBSNs, we define several types of noisy data, and propose the corresponding noisy data filtering techniques.

8. `https://www.facebook.com/`
9. `https://dev.twitter.com/streaming/public`

♢ In **Chapter 4**, from individual perspective, we study user preference on POIs with different granularity and its applications on personalized location based services. Intuitively, users' activities massively imply their preference on POIs. For example, check-ins on restaurants can be regarded as "foot-voting", which means that the user frequently visited restaurants are probably their favorite ones; tips left on restaurant by users often explicitly express their preference about the restaurant and the associated aspects, such as food quality, pricing and environment, etc. In order to study user preference on POIs from these heterogeneous data, we propose a sentiment-enhanced personalized POI recommendation and search framework based on heterogeneous user activity data in LBSNs. First, we define and extract two types of user preference on POIs, namely, *coarse-grained user preference* (i.e., user-POI preference) and *fine-grained user preference* (i.e., user-POI-item preference), from heterogeneous user activity data in LBSNs (e.g., check-ins and user's tips). Afterwards, in order to enable effective personalized POI recommendation and search applications, we develop two novel algorithms based on low-rank approximation techniques for the framework. For the personalized POI recommendation task, we formulate it as a preference prediction problem, and propose a novel location based social matrix factorization algorithm. For the personalized POI search task, we formulate it as a ranking prediction problem, and propose a novel multi-tuple based ranking tensor factorization algorithm. Based on two urban scale user activity datasets from LBSNs, we experimentally evaluate the proposed framework and algorithms. The results show that the proposed framework can subtly capture individual's preference from heterogeneous user activity data, and deliver high-quality personalized POI recommendation and search services.

♢ In **Chapter 5**, from individual perspective, we study the spatial-temporal regularity of user activities in LBSNs and its applications in activity preference inference tasks. The spatial-temporal regularity of user activities has been widely studied. However, different from the traditional studies that are usually based on continuously sampled user trajectory such as in [131], check-ins in LBSNs are user voluntarily reported activities, which usually suffer from a data sparsity problem, causing difficulties in studying the spatial-temporal regularity. Aiming at studying the spatial-temporal regularity of user check-in activities, we propose a spatial temporal activity preference (STAP) model. It first models the spatial and temporal activity regularity separately, and then combine them for activity preference inference. For spatial patterns, we propose the notion of Personal Functional Region (PFR) to model and infer user spatial activity preference. For temporal patterns, we propose to exploit the temporal activity similarity among users and apply non-negative tensor factorization to collaboratively infer temporal activity preference. Finally, we put forward

a context-aware fusion framework to combine the spatial and temporal models for activity preference inference. We experimentally evaluate the proposed STAP model on three urban-scale check-in datasets from LBSNs. The results show that STAP can efficiently model individual spatial-temporal activity preference with sparse check-in data, and consistently outperforms the state-of-the-art approaches in the activity preference inference task.

♢ In **Chapter 6**, from collective perspective, we explore global-scale nation-wide collective activities in LBSNs. Specifically, we design and develop NationTelescope, a platform that monitors, compares, and visualize large-scale collective behavior in LBSNs. First, it continuously collects user behavior data from LBSNs. Second, it automatically generates behavior data summary and integrates an interactive map interface for data visualization. Third, in order to compare and visualize the behavioral differences across countries, it detects the discriminative activities according to the related traffic patterns in different countries. By implementing a prototype of NationTelescope platform, we evaluate its effectiveness and usability via two case studies and a System Usability Scale survey. The results show that the platform cannot only efficiently capture, compare and visualize nation-wide collective behavior, but also achieve good usability and user experience.

♢ In **Chapter 7**, from collective perspective, we explore global-scale city-wide collective activities in LBSNs and their correlation with various cultural factors, such as geography, immigration and religion, etc. We propose a participatory cultural mapping approach to cluster cities into cultural clusters and plot a world cultural map with city granularity. Specifically, since only local users are eligible for cultural mapping, we propose a progressive "home" location identification method to filter out ineligible users. Third, by extracting three key cultural features from daily activity, mobility and linguistic perspectives respectively, we propose a cultural clustering method based on spectral clustering techniques to discover cultural clusters. Finally, we visualize the cultural clusters on the world map. Based on a global-scale user check-in dataset, we experimentally validate our approach by conducting both qualitative and quantitative analysis on the generated cultural maps. The results show that our approach can efficiently capture cultural features from user activities in LBSNs, and generate representative cultural maps. Comparing our cultural maps with those created by traditional cultural mapping approaches based on psychological survey data, we observe not only important cultural correlations between them, but also interesting differences caused by some unique cultural features extracted from user behavioral data.

To conclude, in **Chapter 8**, we discuss and summarize the insights offered by this dissertation. We also present potential directions for future research, such as privacy issues of such location-centric social media data, combination of human activity data and various ubiquitous sensor data, and more innovative applications in smart city scenarios.

1.3 Publication List

During my Ph.D. studies, I have involved in many fruitful collaborations that have yielded 17 publications that span the areas of human dynamics analysis, mobile social media analytics, personalized location based services, context-aware intelligent systems, etc.

Publications related to this dissertation

— **Dingqi Yang**, Daqing Zhang. Participatory Cultural Mapping Based on Collective Behavior in LBSNs. ACM Trans. on Intelligent Systems and Technology (TIST). (under review)

— **Dingqi Yang**, Daqing Zhang, Longbiao Chen. NationTelescope: Monitoring and Visualizing Large-Scale Collective Behavior in LBSNs. Journal of Network and Computer Applications (JNCA). (under review)

— **Dingqi Yang**, Daqing Zhang, Zhiyong Yu, Zhiwen Yu, Djamal Zeghlache. SESAME: Mining User Digital Footprints for Fine-Grained Preference-Aware Social Media Search. ACM Trans. on Internet Technology (TOIT), 2014.

— **Dingqi Yang**, Daqing Zhang, Vincent W. Zheng, Zhiyong Yu. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Trans. on Systems, Man, and Cybernetics: Systems (TSMC), 2014.

— **Dingqi Yang**, Daqing Zhang, Zhiyong Yu and Zhiwen Yu, Fine-Grained Preference-Aware Location Search Leveraging Crowdsourced Digital Footprints from LBSNs. In Proceeding of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013), Zurich, Switzerland, September 2013.

— **Dingqi Yang**, Daqing Zhang, Zhiyong Yu and Zhu Wang, A Sentiment-enhanced Personalized Location Recommendation System. In Proceeding of the 24th ACM Conference on Hypertext and Social Media (HT 2013), Paris, France, May 2013.

Other publications

— **Dingqi Yang**, Daqing Zhang, Korbinian Frank, Patrick Robertson, Edel Jennings, Mark Roddy, Michael Lichtenstern. Providing Real-Time Assistance in Disaster Relief by Leveraging Crowdsourcing Power, Springer/ACM Journal of Personal and Ubiquitous Computing (PUC), 2014.

— Bin Guo, Daqing Zhang, **Dingqi Yang**, Zhiwen Yu, Xingshe Zhou. Enhancing Memory Recall via an Intelligent Social Contact Management System, IEEE Trans. on Human-Machine Systems (THMS), 2014.

— Zhiyong Yu, Daqing Zhang, Zhiwen Yu, **Dingqi Yang**. Participant Selection for Offline Event Marketing Leveraging Location Based Social Networks. IEEE Trans. on Systems, Man, and Cybernetics: Systems (TSMC), 2014.

— Zhu Wang, Daqing Zhang, Xingshe Zhou, **Dingqi Yang**, Zhiyong Yu, Zhiwen Yu. Discovering and Profiling Overlapping Communities in Location Based Social Networks. IEEE Trans. on Systems, Man, and Cybernetics: Systems (TSMC), 2013.

— Zhu Wang, Xingshe Zhou, Daqing Zhang, **Dingqi Yang**, Zhiyong Yu. Cross-Domain Community Detection in Heterogeneous Social Networks. Springer/ACM Journal of Personal and Ubiquitous Computing (PUC), 2013

— Zhiyong Yu, Daqing Zhang, **Dingqi Yang**. Where is the Largest Market: Ranking Areas by Popularity from Location Based Social Networks. In Proceeding of the 10th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC 2013), Italy, 2013.

— Zhu Wang, Daqing Zhang, **Dingqi Yang**, Zhiyong Yu and Xingshe Zhou, Detecting Overlapping Communities in Location-Based Social Networks. In Proceeding of the 4th International Conference on Social Informatics (SocInfo 2012), Lausanne, Switzerland, December 2012.

— Zhiyong Yu, Daqing Zhang, **Dingqi Yang**, Guolong Chen. Selecting the Best Solvers: Toward Community based Crowdsourcing for Disaster Management. In Proceeding of the 2012 IEEE Asia-Pacific Services Computing Conference (APSCC 2012), Guilin, China, December 2012.

— Zhu Wang, Daqing Zhang, **Dingqi Yang**, Zhiyong Yu, Xingshe Zhou, and Zhiwen Yu. Investigating City Characteristics based on Community Profiling in LBSNs, In Proceeding of the 2nd International Conference on Social Computing and its Applications (SCA 2012), Xiangtan, China, 2012

— **Dingqi Yang**, Bin Guo, Daqing Zhang, Better Organizing Your Contacts: An Empirical Study of an Intelligent Social Contact Management System, In Proceeding of the 4th IEEE International Conference on Cyber, Physical and Social Computing (CPSCom 2011), Dalian, China, 2011.

— Bin Guo, Daqing Zhang, **Dingqi Yang**, "Read" More from Business Cards: Toward a Smart Social Contact Management System, In Proceeding of the 10th IEEE/ACM International Conference on Web Intelligence, Lyon, France, 2011.

Chapter 2

Human Dynamics in Social Media: A Literature Review

Contents

Human Dynamics is a transdisciplinary research field focusing on the understanding of dynamic patterns, relationships, narratives, changes, and transitions of human activities, behaviors, and communications [1]. In this chapter, by focusing on the user behavior captured by various social media (particularly the location centric social media), we survey the related work on studying human dynamics from both individual and collective perspectives.

2.1 Human Dynamics from Individual Perspective

The diversity of personality leads to the diversity of user lifestyles, such as food habits and shopping preference, etc. Individual's behavior in social media widely implies such lifestyles of users, i.e., user preference on various items, such as food, music, movies and POIs, etc. The primary research direction in this field is to understand individual users' preference by investigating into their historical activities in social media, in order to perform user preference prediction. Taking food habits as an example, by studying a user's

restaurant visiting records, we can understand what kind of restaurants does the user like, and then predict whether she will like a specific restaurant (or even a specific dash there) or not. Furthermore, in LBSNs, user behavior data usually contains spatial and temporal information, which provide us an opportunity to study the spatial temporal patterns of user activities. In the following, we first summarize the related work on exploring user preference and its applications on personalized information retrieval, and then present the related works on exploring the spatial and temporal patterns of user activities. Finally, by focusing individual human dynamics in location centric social media, we discuss its application on personalized location based services.

2.1.1 Exploring User Preference in Social Media

User behavior in social media massively implies user preference on different types of items, such as music [30], movies [67] and places [155], etc. Given users' historical behavior data in social media, such as music listening records, movie commenting records, and Web page visiting records, etc., we are able to extract user preference on these items and enable personalized information retrieval services, such as personalized recommendation and search. In the following, we present the related work on exploring user preference in social media in two main application scenarios, i.e., personalized recommendation systems and personalized search systems.

2.1.1.1 Personalized Recommendation Systems

Recommendation systems try to suggest users interesting items, such as music and movies. They have been extensively studied in recent years and widely adopted in various commercial web services, such as Amazon [1] and MovieLens [2]. In a typical recommendation system, users' historical behavior data mainly includes their ratings on items, which usually range from 1 to 5, indicating how much the users like the specific items. Based on these historical ratings, recommendation systems try to predict user preference on the unrated items, and thus make the recommendation to users. The underlying intuition of recommendation is that users who share similar preference on some items probably have the similar preference on others.

In academia, a wide range of research work has been done in building recommendation systems using data mining techniques [2]. They mainly fall into three categories: memory-based approach, model-based approach and hybrid approach. Memory-based approaches explore historical rating records to predict unknown ratings without learning step, e.g., classical collaborative filtering methods [118]. They focus on user-item rating matrix and

1. http://www.amazon.com/
2. www.movielens.org/

attempt different strategies to estimate missing ratings. Model-based approaches use the learned model from historical data to predict unknown ratings [94]. They leverage statistics and machine learning techniques to learn models from data in order to predict the missing ratings. Hybrid approaches combine the two aforementioned approaches with certain fusion criterion [2].

The recent growth of social media provides rich social information which can be deployed in recommendation. Unlike traditional recommendation systems assuming that users and items are independent from each other, recommendation systems in social media are able to take the social factor into account. The basic assumption is that users' preference is partially influenced by their social circles. For example, users often resort to their friends or someone they trust for recommendation. According to social relationship type, social network can be divided into two categories: unidirectional and bidirectional. In unidirectional social networks, users establish the relationship without the need of confirmation from others. One example is the follower and following relationship in Twitter. Recommendation based on unidirectional social network can be called trust-based social recommendation [15,60,61,86,91,105]. In bidirectional social networks, the friend relationship can be established if and only if both sides accept it, such as friendship on Facebook. Recommendation based on bidirectional social network can be seen as friend-based social recommendation [62, 87]. While these existing works focus on social recommendation by considering user social network, we believe that considering item similarity can also improve recommendation performance. Hence, we extended the classical matrix factorization approach by considering both user social influence and inter-item similarity in recommendation [149].

2.1.1.2 Personalized Search Systems

With the rapid growth of online information, personalized search has been widely studied in recent years. Some commercial Web search engines, such as Google [3] and Bing [4], have already provided users with personalized search features. Personalized search mainly employs user specific information, such as user context and user preference, to provide users with customized search results. Specifically, it tries to put user designed results on the top of the return list. Based on the type of user information incorporated, personalized search can be roughly classified into two categories: context-aware search and preference-aware search.

Context-aware search leverages user context, e.g., time, location, weather and user activities, to deliver the appropriate search results. Taking local search as an example, a query

3. http://www.google.com/
4. http://www.bing.com

of searching for a burger restaurant for dinner may be interpreted as finding the nearest burger restaurant from one's current location. Maekawa et al. [90] built a context-aware Web search system by incorporating a user's daily activities monitored from ubiquitous sensors. Hansen et al. [54] proposed a general platform to support the development of context-aware hypermedia systems with special emphasis on location-based services, where context-aware search is a main feature. Lane et al. [74] proposed a context-based local search framework that considered rich context such as weather and activity. Iwata et al. [59] extracted user's situation, e.g., being in the office at lunch time on weekdays, or going downtown on holiday, to perform personalized search. Since context-aware search mainly incorporates users' current context of submitting queries to generate customized search results, it does not practically handle individual's personal preference, which plays an important role in delivering personalized search results. In the following, we focus on leveraging user preference for search personalization, i.e., preference-aware search.

Preference-aware search provides search results according to individual's preference. It has been widely studied in Web search personalization. Eirinaki et al. [45] conducted a survey of Web personalization using user preference. There are roughly two ways of obtaining user preference. The first approach leverages user explicit feedback, i.e., let user explicitly state their preference on search results. However, according to an early user study [5], users usually do not want to spend extra efforts providing such information. The second approach uses implicit feedback. Since it can be collected without extra user efforts, they are widely used in search personalization to extract user preference. The classical implicit feedback sources include browsing history [132], click-through data [97] and user personal information (e.g., email, desktop data) [34].

The booming of social networks brings a new opportunity for collecting user feedback to enable personalized search, such as image search in Flickr [5] [123], scholar search in CiteULike [6] [65], web bookmark search in Delicious [7] [21], etc. In most of social network services, users can add tags and make comments on social media items (e.g., photos, video, blogs, POIs). Such data can be regarded as user direct feedback and massively implies their preference, which can then be used in search personalization [25, 123, 147, 164]. According to how user preference is used in the personalization schemes, preference-aware search can be classified into three categories.

First, user preference can be used to augment user submitted query with keywords (i.e., query expansion). For example, using crowdsourcing data from the social bookmarking web service Delicious, Zhou et al. [164] extended original query using user profile extracted

5. `http://www.flickr.com/`
6. http://www.citeulike.org/
7. `http://www.delicious.com/`

from one's social annotation history. The approaches in this category usually have limited personalization capability due to the lack of user preference in the ranking process.

Second, user preference can be used to re-rank the search results generated from a non-personalized search engine. For example, Xu et al. [147] extracted users' interests from social annotation data and ranked documents according to both query-document relevance and similarity between users' interests and documents' topics. By constructing user profiles and resource profiles, Cai et al. [25] first modeled query-document relevance and then leveraged user-document preference to adjust the result ranking. These studies separate query-document relevance ranking and user-document preference ranking, and then merge them together.

Third, user preference can be used in the document indexing and searching process. The most popular approach supporting this scheme is tensor factorization. In Web search, Sun et al. [133] conducted an early work by modeling click-through data as a three-way tensor and then using High-Order Singular Value Decomposition (HOSVD) techniques to factorize the built tensor for personalized ranking. Sang et al. [123] proposed a multi-correlation ranking approach in tensor along with a user-specific topic modeling to personalize image search using social annotation data on Flickr. Since the existing tensor factorization methods cannot handle the multi-tuple ranking problem, we proposed a novel multi-tuple based ranking tensor factorization algorithm to perform personalized ranking in local search scenario [150, 151].

2.1.2 Exploring Spatial Temporal User Activity Patterns in Social Media

In location centric social media, users generate a considerable volume of spatial-temporal activity data. Using these user behavior data, we can study user mobility, and try to predict the future/next locations where a user will be. Moreover, since the user activities in LBSNs are usually recorded with the activity semantics. For example, a user's check-in at a French restaurant probably means that the user is having French food there. Combining such semantics with user mobility, we can study the spatial temporal regularity of user activities and infer one's current activity semantics. In the following, we first summarize user mobility related works, and then present the existing work on studying spatial temporal user activity semantics.

2.1.2.1 Human Mobility

The study on human mobility can be dated from 1885, when Ravenstein published his work on studying migration [114]. He summarized three properties which are still applicable in the modern world, i.e., most migration is over a short distance; long range migrants usually move to urban areas; migration increases with economic development. In the early

1990, with the emergence of cellular communication technology, users left their mobility traces when carrying their mobile phone in the daily lives. It was the first instance in human history that human mobility can be tracked in real-time with relatively high geographical precision. With such data, Gonzalez et al. [50] studied individual human mobility patterns and demonstrated the high degree of temporal and spatial regularity of human trajectories. Recently, the GPS-embedded smartphones provide a novel opportunity to obtain precise location (i.e., GPS coordinates) of users. From 2009 to 2011, Nokia Research Center, Idiap Research Institute, and EPFL (Swiss Federal Institute of Technology in Lausanne) conducted a user mobility data collection campaign in Lausanne and its surrounding areas, which included 200 participants [75]. Using this dataset, Do et al. [43] studied human mobility and proposed a contextual conditional human mobility model. However, due to the privacy issue, it is practically difficult to conduct such data collection campaign on a large population.

The recent rise of location centric social media presents a novel source of user mobility data. By carrying GPS-equipped smartphones, users voluntarily share their precise location in LBSNs. With the increasing popularity of LBSNs, large-scale user mobility data becomes attainable. Different from continuously sampled user mobility traces [75], due to the fact that most of users do not regularly and frequently perform check-ins, user check-in data in LBSNs is sparse. Although such sparsity causes difficulties in modeling human mobility, in current literature, various research works prove that user check-ins in LBSNs still show obvious mobility pattern. For example, by investigating into user check-in data in LBSNs, Noulas et al. studied the spatial temporal patterns of user mobility [103], and showed the existence of a universal power-law distribution in the physical distance of human movement [100]. By collecting user check-in data in the United States, Cheng et al. [33] reported a quantitative assessment of human mobility patterns by analyzing the spatial, temporal, social, and textual aspects. Cho et al. [35] studied the social relationship and user mobility patterns, and built a mobility model based on their findings. They showed that periodical behavior can explain the majority (50%-70%) of user movement, while social relationship can explain a part (10%-30%) of user movement.

By studying user mobility, the most straightforward application is mobility prediction, which can enable various applications. For example, knowing a user will go back home soon, the heating system in the user's house can be started before the arrival of the user, in order to improve the user's comfort with optimized energy consumption. In LBSNs, location prediction in terms of POIs aims at predicting the specific POI that users will visit next. For example, Chang et al. [29] incorporated various features in LDA model for next POI prediction. Gao et al. [47] proposed a social-historical model based on Hierarchical Pitman-Yor process for predicting the next check-in of a user. Noulas et al. [101] extracted user

specific features and global mobility features and built a next POI prediction model.

2.1.2.2 Spatial Temporal Activity Semantics

Different from the classical user mobility data that only contains plain GPS readings, user check-in data in LBSNs usually contains rich information. Specifically, in current literature [80, 99, 111, 153], the POI categories of check-ins can be considered as the semantic interpretation of user activities (e.g., food, shopping, entertainment). With such information, we cannot only understand the plain user mobility, but also investigate into their activity semantics. For instance, Lian et al. [80] clustered users based on their temporal activities and activity transition in order to collaboratively identify user activities. Pianese et al. [111] clustered user activities for user routine detection and predicted user future activities as well as locations. Ye et al. [153] used the mixed hidden Markov model to predict user's next activity. Based on user activity data in LBSNs and cellular data from telecommunication providers, Noulas et al. [99] studied the semantic activity inference problem in urban areas. By studying the spatial temporal regularity of user activities, we proposed a novel spatial temporal activity preference model to infer individual activity preference given a user's current context, i.e., location and time [152].

2.1.3 Applications of Individual Human Dynamics in LBSNs

By exploring individual human dynamics in LBSNs, we are able to build various personalized location based services. Typical location based services include location search [74], location recommendation [149], etc.

Location recommendation tries to suggest users interesting places to visit. Existing location recommendation can be divided into two categories: 1) generic location recommendation and 2) personalized location recommendation. First, generic location recommendation usually provide users the most popular venues according to public opinions such as in [26]. Due to the lack of individual preference, users receive identical recommendation from such systems. Second, personalized location recommendation aims at providing users with the most pertinent venues by considering individual's preference. Among various personalized location recommendation approaches such as classical collaborative filtering [11], matrix factorization [19, 31, 154, 155] and recommendation with random walk [102], matrix factorization is the most popular approach due to its online recommendation efficiency. Different from using user-item rating records in classical matrix factorization approaches, location recommendation in LBSNs mainly takes user's check-ins as inputs. The most popularly used model is 0/1 scheme, i.e. the places users visited are labeled as 1 and non-visited as 0. Using this model, Ye et al. [154, 155] studied the geographical and social influence in point-of-interest recommendation based on collaborative filtering techniques. Another

model is based on check-in frequency which quantifies users' preference on venues according to the number of their check-ins. With this scheme, Berjani et al. [19] developed a location recommendation system using matrix factorization methods. Chen et al. [31] proposed a multi-center Gaussian model to capture the geographical influence and combined the matrix factorization with social regularization to perform the location recommendation. All of these location recommendation systems use the user check-in information to model user preference. Aiming at improving the effectiveness of location recommendation, we proposed a hybrid user POI preference model by combining the preference extracted from check-ins and text-based tips which were processed using sentiment analysis techniques [149].

Personalized location search mainly employs user specific information such as user context and user preference, to provide customized search results. Most of the existing personalized location search approaches exploit user context. For example, with consideration of user's current location, Choi [36] utilized fuzzy query techniques to re-rank the search results. Leveraging user's current location and time, Waga et al. [140] built a location search system using context-aware recommendation techniques. By studying the spatial temporal patterns of user activity, Iwata et al. [59] extracted user's situation, e.g., being in the office at lunch time on weekdays, or going downtown on holiday, to perform personalized search. Lane et al. [74] proposed a framework that considered rich context such as weather and activity. While there are few studies on preference-aware location search, we proposed a fine-grained preference-aware location search framework [150], which leveraged heterogeneous user feedback on POIs (i.e., check-ins and text-based tips) to extract user preference with finer granularity, and incorporated such fine-grained user preference in personalized ranking process using tensor factorization techniques.

2.2 Human Dynamics from Collective Perspective

Collective behavior has been widely studied in the long history of human development. Turner et al. [138] define collective behavior as the behavior of aggregates whose interaction is affected by some sense that they constitute a group but who do not have procedures for selecting or identifying leaders or members. For example, French people often go to French restaurants in the evening for dinner while Japanese usually go to bars after work. In the following, we first summarize the related work on studying collective behavior in social media, and then discuss its applications.

2.2.1 Collective Behavior Analysis in Social Media

Social media presents a rich source for studying collective behavior. According to the activity types on social media, the collective behavior can be classified into two categories,

cyber behavior and physical behavior.

2.2.1.1 Collective Cyber Behavior

When users interact with each other via online social media, their cyber behavior (such as messaging, clicking, web-page visiting, etc.) is recorded by these social media services. By studying such data, we can explore large-scale online crowd activities. For example, Benevenuto et al. [17] conducted an analysis on user cyber behavior in various online social network services. They studied how users interact with their friends (e.g., frequency of communication, online activity sequence, etc.) in social networks. By studying the collective behavior in Flickr social network, Cha et al. [28] studied the information propagation patterns. Golder et al. [49] studied the cultural differences of the collective moods by applying sentiment analysis on user messages in Twitter. Park et al. [108] investigated the differences on the usage of facial expressions for emotion in Twitter in different countries. While these works shed light on the underlying patterns of collective behaviors, they are limited on the cyber activities in the virtual world of the Internet.

2.2.1.2 Collective Physical Behavior

With the advent of location centric social media, users can share their physical activities (such as having dinner in a restaurant) within their social circle. Such physical behavior has recently gained increasing popularity in studying collective behaviors. For example, Preotiuc-Pietro et al. [113] explored collective behavior based city-to-city similarity measures, which considered a city as a bag of user activity categories. Based on user physical activity data in LBSNs, Noulas et al. [104] clustered and annotated regions in a city. Using collective user activity in LBSNs, Wang et al. investigated the overlapping community detection problem [141, 143], and further studied the cross-domain community detection problem in heterogeneous social networks [144]. Cheng et al. [33] studied collective user mobility patterns in the US with regard to geographical, economic and social factors. Bauer et al. [14] discovered the dominant topics in the neighborhoods of a city by applying topic modeling techniques on collective user activity data. Wang et al. [142] investigated city characteristics based on community profiling in LBSNs. Yuan et al. [160] explored both individual and community lifestyles in China using user digital footprints left in LBSNs and other social networks. Silva et al. [129] studied large-scale city dynamics and identified several cultural differences on eating habits across different cities. Although these works provide insight into the characteristics and regularities of user collective behavior in LBSNs, they are usually limited by the collected datasets, i.e., fixed datasets with a small or moderate scale (e.g., check-in data in a city or a country during several weeks or months). Aiming at studying the large-scale collective behavior, we introduce the NationTelescope

platform in Chapter 6 to collect, analyze and visualize the user check-in behavior in LBSNs on a global scale. Moreover, by studying the correlation between global cultures and various aspects of collective activities in LBSNs, in Chapter 7, we propose a cultural mapping approach to discover and visualize the global cultures from city-wide collective behavior.

2.2.2 Applications of Collective Human Dynamics in LBSNs

Understanding collective human dynamics can enable various applications. In the following, we briefly survey the two main research directions, i.e., event detection and urban planning.

Event detection in social media has been extensively studying in recent years. When users interact with each other via social media, they can be regarded as "social sensors". Social media thus collects massive such human sensing data, which can be used in event detection [145]. A notable work from Sakaki et al. [120] studied user collective behavior in Twitter, and its application on event detection of earthquake in Japan. Cataldi et al. [27] studied the emerging topic detection problem on Twitter. Based on user check-in data in LBSNs, Liang et al. [81] studied the correlation between social event and collective mobility, and propose an event and location based population model.

Urban planning using social media data is an emerging research topic. User activities in social media are mainly located in urban areas, which present a new data source for urban monitoring and planning. From urban environment perspective, Zheng et al., studied the air-quality inference problem [162] and urban noise categorization problem [163] by combining heterogeneous data sources, such as meteorology, traffic flow, human mobility, structure of road networks, and point of interests. From urban economic perspective, Karamshuk et al. [63] studied the retail store placement problem using user check-in data in LBSNs. They investigated into the influence of various factors (e.g., collective mobility in cities, store location and its surrounding) on retail business to identify the optimal store placement. Yu et al. studied the market effects of different types of business (e.g., food, entertainment, shop, nightlife, etc.) based on user check-in data [156] in LBSNs and further the participant selection problem for off-line event marketing [158].

Chapter 3

Human Activity Data Collection From Location-Centric Social Media

Contents

3.1 Data Collection Platform

The soaring popularity of location based social networks makes large-scale human behavior data become attainable. In order to obtain big user activity data from LBSNs, we design and develop a social media data collection platform, which can continuously collect real-time user activity data from various LBSNs in a streaming manner. This platform is developed within an European FP7 project SOCIETIES [1]. In the following, we first present the overview of SOCIETIES project, and then present the data collection platform, followed by the detailed data collection process.

1. http://www.ict-societies.eu/

3.1.1 SOCIETIES Project Overview

The SOCIETIES (Self Orchestrating CommunIty ambiEnT IntelligEnce Spaces, No. 257493) project aims to investigate and address the gap between pervasive and social computing by designing, implementing and evaluating an open scalable service architecture and platform for pervasive communities. A pervasive community is inherently context-aware, self-organizing, self-improving and capable of pro-active behavior aiming to optimize and personalize the pervasive experience of an entire community. In addition to the resources controlled by its individual members, a pervasive community may also provide public access to its devices, services and resources. The notion of Cooperating Smart Spaces (CSSs) has been introduced so as to extend pervasive systems beyond the individual to dynamic communities of users. SOCIETIES project enables the *Discovery*, *Connection* and *Organization* of relevant people, resources and things, crossing the boundary between the real and virtual worlds. The vision of SOCIETIES can be categorized in terms of three broad phases each of which contributes to the formation of our Cooperating Smart Spaces which are: Discover, Connect and Organize. The use cases of SOCIETIES project include student community discovery on campus, professional community organization in enterprise scenarios, and intelligent disaster management. For example, the concept of the pervasive community can be used to build community based disaster management system [157], which can provide real-time assistance in disaster relief by leveraging crowdsourcing power [148].

As an open scalable service platform, SOCIETIES needs to be integrated with the existing online social network services. Users can access different social network services via SOCIETIES platform. To achieve this goal, we develop a social network connector within SOCIETIES project. It provides a proxy between SOCIETIES platform and user social communities in the existing online social network services. It cannot only connect to one or more social networks and fetch most of the user profile information, and their activities, but also provides an Application Programming Interface (API) to push data through one or more social channels. In this chapter, since we focus on the human activity data collection from LBSNs, we only present the data collection part of the social network connector in SOCIETIES platform.

3.1.2 Platform Design and Component

Figure 3.1 illustrates the architecture of the proposed data collection platform. It is composed of several LBSN Connectors, a Social Data collector, an Access Controller and an Access Token Database. The consideration of such a design is mainly due to the access control scheme of individual LBSNs. Specifically, due to the privacy protection of

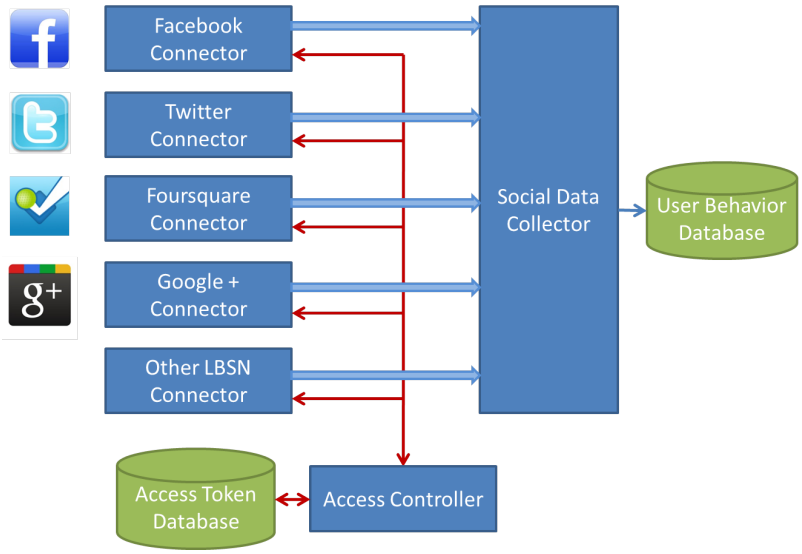


Figure 3.1 – Data collection platform.

users' personal data, most LBSNs integrate the OAuth protocol [2], an open standard for authorization. In order to access the data stream from LBSNs, an authentication process is required with specific access tokens. Therefore, we implement the Access Controller and Access Token Database for authentication with various LBSNs. In addition, since different LBSNs usually provide different APIs, the corresponding LBSN connector is implemented under the specification of each LBSN.

After the authentication with LBSNs, the Social Data Collector component continuously gathers user behavior data. In order to handle the data heterogeneity across different LBSNs, we adopt OpenSocial API [3], which is a public specification of social network Web framework supporting an extendable data structure for different social networks. Moreover, such a design of data collector also ensures the scalability of the platform when adding new LBSNs. Specifically, we can easily incorporate new LBSNs in our platform by only implementing the corresponding LBSN connectors.

3.1.3 Data Collection Process

Using this data collection platform, we can collect user activity data from various LBSNs (i.e., Twitter, Foursquare and Facebook). Due to the privacy protection policy in social media services, user activity streams may not be accessed publicly. For example, most users' activity data in Facebook and Foursquare can only be accessed with the permission of the

2. `http://oauth.net/`
3. `http://opensocial.org/`

(a) Foursquare check-in user interface (b) Shortened check-in URL in Tweet (Json format)

Figure 3.2 – An example of Foursquare check-in sent via Twitter.

users (i.e., specific access tokens granted by users). Therefore, large-scale data collection directly from these social media is not feasible, because it is practically difficult to obtain permission from a large number of users.

Fortunately, Twitter public streams [4], which massively contain user activity data from various social media (e.g., Foursquare and Facebook), can be accessible without user permission. Specifically, via Twitter, users can send a Tweet (i.e., a short text) to share their current status, which is publicly visible by default. Due to the popularity of Twitter, some other social media integrates Twitter in their services. For example, in Foursquare, users may share their check-ins as a Tweet. Figure 3.2(a) demonstrates the check-in user interface of Foursquare, where users can share their check-ins via Twitter and Foursquare. In Twitter public streams, such a check-in is identified by a shortened URL linked to the original social media services. For example, Figure 3.2(b) shows the shortened URL of a Foursquare check-in shared in Twitter. By resolving such a URL, we can obtain a full check-in activity in Foursquare. Following the above process, we can continuously collect user activity data from LBSNs in a streaming manner.

4. `https://dev.twitter.com/streaming/public`

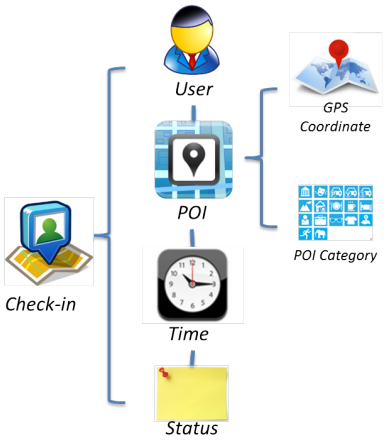


Figure 3.3 – Check-in data format in Foursquare.

3.2 Data Format and Representation

In this dissertation, we mainly collect user activity data from Foursquare, which is one of the most popular LBSNs. Since its creation in 2009, it has attracted more than 45 million users globally and contained more than 5 billion check-ins by January 2014 [5]. In the following, we present the collected user activity data in Foursquare.

First, check-ins are the most important user activity data in Foursquare. A typical check-in represents a user's presence at a POI at a specific time with a short message expressing her current status. Figure 3.3 illustrates the data format of check-ins in Foursquare.

— A user is uniquely identified by a user ID, which is linked to a specific user account.

— A Point of Interest (POI) or venue is uniquely identified by a venue ID. It contains the GPS coordinates of the venue, and its category information, such as a bar or a restaurant.

— A time is the UTC time when a check-in is conducted, such as "Wed Jan 29 14:55:24 +0000 2014". With the GPS coordinates of the checked POI, we can obtain the local time offset, and thus convert the UTC time to the local time of the check-in places.

— A status of a check-in is a short text-based message, which usually expresses the user's current status, such as "Having fun with my buddies here."

— A GPS coordinate is represented in the format of decimal degrees. The latitude is preceded by a minus sign if it is south of the equator (a positive number implies north), and the longitude is preceded by a minus sign if it is west of the prime meridian (a

5. `https://foursquare.com/about`

positive number implies east). For example, the GPS coordinate of a POI in Paris is "48.852954, 2.338000".

— A POI category can be regarded as the semantic representation of the POI, such as a bar or a restaurant, etc. In Foursquare, POIs are organized with a three-level hierarchical category classification [6]. It contains 9 root categories (i.e. Arts & Entertainment, College & University, Food, Great Outdoors, Nightlife Spot, Professional & Other Places, Residence, Shop & Service, Travel & Transport) which are further classified into 291 categories at the second level. In addition, a part of second-level categories are divided into sub-categories at the third level. Due to the fact that Foursquare continuously updates the classification of POI categories, the number of the third level categories varies over time. By the end of January 2014, it contains 437 POI categories at the third level.

Second, there are other POI related data, such as tips and tags. Tips are text-based comments that users left on POIs, which usually express users' opinions about the checked places. According to the post [7], about two thirds of Foursquare users post tips on POIs. Compared to the check-in status which tends to express the real-time personal feeling, tips of a venue are more like customer reviews. For example, a tip left on an Italian restaurant is "Good place in center New York, I went there last Sunday night and had great spaghetti with reasonable price. But I had a very long waiting time, almost one hour just for appetizer!!!" Such tips on POIs not only imply users' feeling about the POI, but also the different items/aspects of POIs (e.g., dishes, environment and pricing, etc.). In addition, users can also add tags on POIs to characterize them. For example, the tags of an Italian restaurant include "salad", "seafood", "spaghetti", "pizza" and "free-wifi", etc. Such tags can be used to label and index POIs.

3.3 Noisy Data Filtering

Noisy data is inevitable in social media. By analyzing the collected user activity data, we identify three types of noisy data.

First, check-ins from users who have ever performed "sudden-move" check-ins are considered as noisy data. Even though Foursquare tries to verify whether a user is actually near the place when she checks in, fake check-in data still exist. For example, in order to get some awards in Foursquare, some malicious users may use Foursquare API to perform fake check-ins. We observe that some users have ever performed "sudden-move" check-ins (consecutive check-ins with a speed faster than 1200 km/h, i.e., movement faster than the

6. https://developer.foursquare.com/docs/venues/categories
7. http://techcrunch.com/2011/08/04/klout-adds-foursquare-but-how-much-will-it-boost-my-score/

common airplane speed). These "sudden-move" users represent about 1.1% of all the users, while their check-ins represent about 3.4% of all the check-ins in the collected data. All the check-ins from these "sudden-move" users are considered as noisy data and thus eliminated.

Second, check-ins at the POIs without venue category are considered as noisy data. Specifically, some of the venues cannot be resolved by Foursquare venue API, causing the venue category information of these venues to be unavailable. These venues present about 7.5% of all the venues, while the check-ins at these venues represent less than 1.0% of all the check-ins in the collected data because these venues are usually unpopular. Since venue category is critical to semantically understand user activity, we thus exclude the check-ins which were performed at these venues.

Third, check-ins of the inactive users are considered to be less representative of the user community of the social media in this thesis, and thus need to be filtered out. In most of social media, some users may be very inactive. We acknowledge that these user activities may be important in other studies, such as the study examining why people use LBSNs [84], but they are not the focus of this thesis. In many social media services, there exist a large number of such inactive accounts. For example, according to Nielsen's "Social Media Report 2012" [8], there are only 42% of Foursquare users who used it at least once a month. The same metric in our dataset is about 39%. In this dissertation, except specifically mentioned, the inactive users are defined as the users who have performed less than one check-in on average per week. According to this definition, the inactive users represent about 86.1% of all the users, while their check-ins represent about 38.2% of the total check-ins. Therefore, we filter out the check-ins of these inactive users in the collected data.

Since we continuously collect user activity data from LBSNs, the volume of the collected dataset continuously increases. To give a reference, from April 2012 to September 2013, we collect 81,571,174 check-ins conducted by 2,418,223 users at 10,428,709 venues globally. After noise filtering, the dataset contains 49,273,956 check-ins conducted by 279,495 users at 6,743,711 venues.

8. http://www.nielsen.com/us/en/insights/reports/2012/state-of-the-media-the-social-media-report-2012.html

Chapter 4

Understanding User Preference on POIs

Contents

4.1 Introduction

User activities in LBSNs massively reflect individual preference, which can thus enable various location based services. In this chapter, we propose a sentiment-enhanced personalized POI recommendation and search framework based on heterogeneous user activity

data in LBSNs. Specifically, we explore two types of user preference on POIs with different granularity (i.e., coarse-grained user preference and fine-grained user preference), and then study their applications in two typical location based services, i.e., location recommendation and search, respectively.

4.1.1 User Preference with Different Granularity

When using LBSNs, users leave heterogeneous digital footprints, such as check-ins and tips, which implies user preference with different granularity. In this dissertation, we define two types of user preference on POIs, viz., coarse-grained user preference and fine-grained user preference.

— On the one hand, *coarse-grained user preference* describes a user's general and board feeling about a specific POI. For example, check-ins on restaurants can be regarded as "foot-voting". Intuitively, users may probably prefer the restaurants where they frequently visit. Such user-POI preference is a typical coarse-grained user preference.

— On the other hand, *fine-grained user preference* describes a user's specific feeling about different items [1] at a POI (i.e., user-POI-item preference). For example, by leaving a tip at a restaurant, a user explicitly express her preference. A tip of a burger restaurant may be "I like cheese burgers at this restaurant, but not the beer there". By applying sentiment analysis on this tip, we can extract her positive preference for "cheese burgers" and negative preference for "beer" at this POI.

By exploring these two types of user preference, we can enable personalized location based services. First, by studying coarse-grained user preference on POIs, we can predict user preference on their unvisited POIs using data mining techniques, and thus achieve recommendation tasks. Second, by extracting fine-grained user preference on POIs, we can predict users' preference on different POIs with regard to a specific item, and thus enable personalized POI search.

4.1.2 Coarse-grained User Preference and POI Recommendation

Different from the classical recommendation systems with explicit rating records which reflect users' preference, POI recommendation in LBSNs usually utilizes user's behavior, i.e. check-in, to model users' coarse-grained preference on POIs [19, 31, 102, 154]. Nevertheless, merely using check-in data has two shortcomings. First, check-in data of a user may not be sufficient to reflect her preference. Compared to web based rating services which capture users' preference on items, check-ins only represent users' habitual behaviors. Intuitively,

1. A "item" here has a board meaning. It may not only refer to a physical entity (e.g., a special dish or drink in a restaurant), but also refer to a specific aspect (e.g., environment and pricing of a restaurant).

users prefer those venues with high check-in frequencies. However, those less checked venues may not be necessarily less favored by users. Second, check-in frequency is directly considered as the degree of users' preference in POI recommendation, the negative feedback in the comments made in each venue is not taken into account, which may introduce biases to the user preference measure. Besides user preference model, recommendation algorithm should also be improved to handle both inter-user and inter-venue relationships. The state-of-the-art POI recommendation approaches only consider how user social network can influence recommendation results [62, 87]. But in fact, POI recommendation needs to consider more factors such as the similarity between POIs.

Aiming at solving the two aforementioned problems in location recommendation, we first propose a novel user preference model with extra information besides check-in and then extend matrix factorization methods in classical social recommendation to capture both social and inter-venue influence.

First, we consider both user check-ins and comments on venues in POI recommendation. While check-in frequency represents how much users prefer POIs, tips need to be further processed in order to extract user preference from them. We use text-based sentiment analysis techniques to extract one's sentiment in tips and then convert it as a measure of coarse-grained user preference. We also propose a fusion framework to get a unified preference model from both check-ins and tips.

Second, venues can construct a similarity network according to their categories. Similar to user social network, we believe that venue similarity can also influence recommendation performance. Therefore, by constructing a user-POI preference matrix, we formulate the POI recommendation problem as a preference prediction problem, and introduce a Location Based Social Matrix Factorization (LBSMF) method to capture the influence on preference prediction from both user social network and venue similarity network perspectives.

4.1.3 Fine-grained User Preference and POI Search

Personalized POI search is usually fulfilled from two perspectives in current literature, i.e., via context-awareness and preference-awareness. The *context-aware search* leverages user's context, e.g., current time, location, weather condition, user's activity, to augment the search queries and deliver the appropriate search results to users. For example, a query of looking for a burger restaurant can be interpreted as finding "where is the *nearest* burger restaurant" (i.e., location context is taken into account). The *preference-aware search* provides results according to the individual's preference about venues. The same query above may be interpreted as asking "where is the burger restaurant serving *my favorite taste of burgers*". Most of the research efforts on location search personalization focus on search according to context. Even though there are less efforts focused on preference-aware

location search, as a special type of information in the web, locations can be retrieved using personalized web search approaches. Web search personalization is an extensively studied topic where user preference has been widely used to enhance user's search experience. In Web search, preference-aware approaches usually provide users with a personalized list of results using certain ranking algorithm by incorporating user preference which is mainly extracted from users' historical search records.

In order to build an effective preference-aware location search services, we develop a fine-grained preference-aware location search framework leveraging user activity data in LBSNs. In particular, we exploit or introduce unique features in three key phases of the preference-aware location search scheme, i.e., in *user feedback capture*, *user preference modeling* and *search result ranking*.

1) *Collecting users' direct feedback on venues from LBSNs.* Users' interaction in LBSNs can be regarded as user feedback on locations. Different from the classical user feedback on web based location search (e.g., click-through data, browsing history, past queries), the user feedback from LBSNs is direct and more precise. For example, a user intends to search a French restaurant in New York, clicking on one restaurant's website does not indicate that she would go to that place. Even if she goes to the restaurant later, this might not necessarily mean that she would like the restaurant. However, in LBSNs, users physically visit and leave comments directly on the venues. Thus collecting the user feedback would better characterize users' actual feeling about the venues and what entities users like/dislike at those venues.

2) *Modeling fine-grained user preference extracted from heterogeneous user feedback in LBSNs.* User generated traces at venues in LBSNs are usually heterogeneous, including check-ins, tags in terms of keywords, and tips in the form of short text. These contents imply user preference at different granularity levels, i.e., coarse-grained user preference and fine-grained user preference. Apparently, compared to coarse-grained user preference (in the form of user-POI), fine-grained user preference (in the form of user-POI-item) contains more precise and detailed information. It provides us with new possibilities to rank locations more accurately according to one's preference.

3) *Incorporating fine-grained user preference into personalized location ranking using tensor factorization techniques.* As discussed in Chapter 2, in order to incorporate such fine-grained user preference in the indexing and searching process, the most popular approach is based on tensor factorization. Concretely, a three-way tensor is adopted to model the fine-grained user preference. In web search, Sun et al. [133] conducted an early work by modeling click-through data as a three-way tensor and then using High Order Singular Value Decomposition (HOSVD) techniques to factorize the built tensor for personalized ranking. Since HOSVD cannot practically handle sparse tensors, Rendle et al. [115] pro-

posed a ranking with tensor factorization approach to specifically address element ranking problem in tensor, which can alleviate the problem of sparsity. Sang et al. [123] proposed a multi-correlation ranking approach in tensor along with a user-specific topic modeling to personalize image search using social annotation data on Flickr. Since ranking tensor factorization process is usually time-consuming, Rendle et al. [117] designed a pairwise interaction tensor factorization model which dramatically reduced the learning time while maintaining the performance. Shi et al. [127] addressed the top-N context-aware recommendation problem by leveraging tensor factorization to maximize mean average precision. However, existing tensor factorization approaches can merely handle positive preference. As fine-grained preference might include both positive and negative preference, we propose Multi-Tuple based Ranking Tensor Factorization (MT-RTF) to consider both positive and negative preference simultaneously in the factorization process [150].

4.2 Framework Overview

Figure 4.1 illustrates the proposed sentiment-enhanced personalized POI recommendation and search framework. The left panel shows users' activities in physical world while the right part presents the framework components. The data collector is in charge of gathering raw data from LBSNs. The user preference analysis component extracts coarse-grained and fine-grained user preference information from heterogeneous user activity data using the corresponding techniques. Concretely, we use statistical analysis technique to process check-in and tag data, and use text-based sentiment analysis technique to process tip data. Afterwards, for personalized POI recommendation, we formulate it as a preference prediction problem, and leverage the proposed LBSMF algorithm to predict the missing user preference. For personalized POI search, we formulate it as a ranking prediction problem, and use the proposed MT-RTF algorithm to achieve personalized ranking.

4.3 User Preference Analysis

The heterogeneous user activity data in LBSNs implies user preference with different granularity. In this section, we extract user preference from both check-in and tip data, and convert them to coarse-grained and fine-grained user preference. Specifically, for coarse-grained user preference, we propose a Hybrid Preference Model (HPM) unifying user's preference in both check-ins and tips to build a user-venue preference matrix. For fine-grained user preference, we propose a tensor based user preference model that first augments user preference from check-ins with the help of tag data, and then merge it with fine-grained user preference from tips.

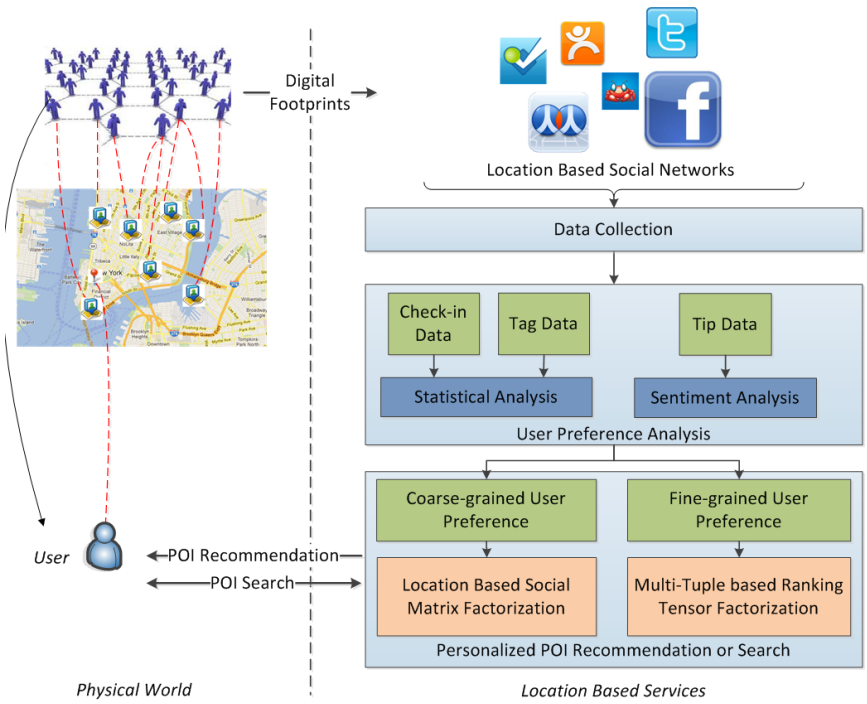


Figure 4.1 – Sentiment-enhanced personalized POI recommendation and search framework.

4.3.1 Coarse-gained User Preference Modeling

In order to extract coarse-grained user preference from check-ins and tips, we propose a Hybrid Preference Model that first analyze user preference from individual data, i.e., check-ins and tips, and then propose a fusion framework to build a unified user-venue preference matrix.

4.3.1.1 User Preference from Check-in Data

Based on the number of check-ins, a user-venue preference matrix can be built. Without loss of generality, we use a five-point preference scale in the preference matrix, where 1 represents for "Poor", 2 for "Fair", 3 for "Good", 4 for "Very Good" and 5 for "Excellent". Due to the power law distribution of user-venue check-in numbers [37], the number of check-ins is mapped as follows: one check-in corresponds to 2, two check-ins to 3, three check-ins to 4, and four or more check-ins to 5, resulting in a *check-in preference matrix*.

4.3.1.2 User Preference from Tip Data

Tips are short texts that often describe users' comments about venues, which can be processed using sentiment analysis techniques. In this work, the dictionary based unsuper-

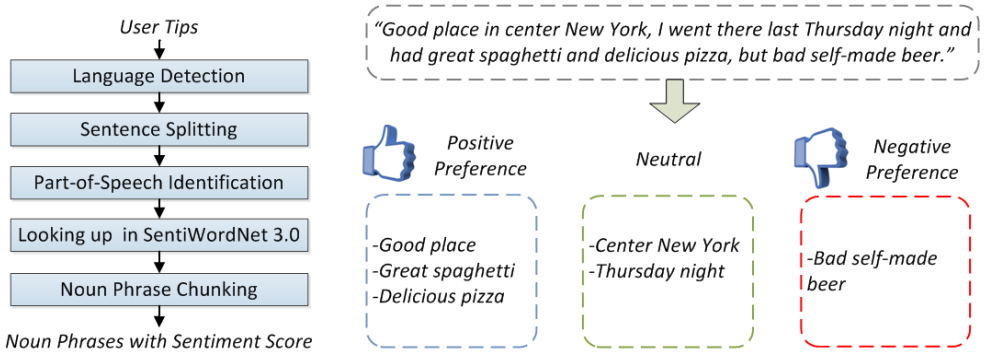
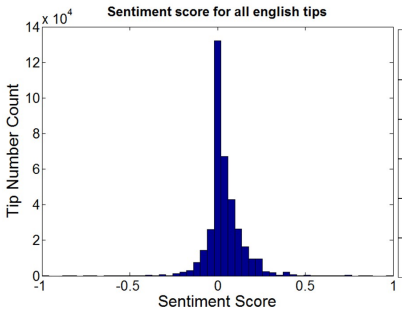


Figure 4.2 – Sentiment analysis of tips.

vised sentiment analysis method is used for the sake of its simplicity in implementation. More sophisticated sentiment analysis techniques can also be applied to improve the performance, but they are not the focus of our work. Figure 4.2 demonstrates the sentiment analysis process of tips. The left part shows the processing workflow. The right part illustrates the sentiment analysis results of an example tip about an Italian restaurant.

In our study, we merely deal with tips in English. The language detection component firstly filters out non-English tips. We use a language detection library developed by Cybozu Labs [128]. Then tips are split into sentences and identified the part-of-speech for each word, e.g., "good" is an adjective, "place" is a noun, "went" is a verb. Afterwards, we can obtain a sentiment score for each word referring to SentiWordNet [9] with the corresponding part-of-speech type. The positive, zero and negative values of the sentiment score indicate the positive, neutral and negative sentiment, respectively. Noun-Phrase Chunking is then performed to get the phrases e.g., "good place", "delicious pizza", which describe what users like or dislike at a venue. The overall sentiment score of a tip is the sum of all the sentiment scores of each word in the tip and is normalized into [-1, 1], where -1 and 1 represent the most negative and positive sentiment, respectively. The implementation is based on NTLK toolkit [85].

Given the overall sentiment score of a tip, we need to map it to the user preference score ranging from 1 to 5. The mapping scheme should also consider its statistical distribution. As shown in the left part of Figure 4.3, the distribution of sentiment scores is highly centralized around 0, i.e. neutral sentiment. This implies that most of the tips have the sentiment around neutral. Furthermore, a slight bias towards positive sentiment is also observed, which implies people tend to leave more positive tips at the venues where they checked in. Considering such a distribution of sentiment scores, we propose a mapping scheme for sentiment scores (presented in the right part of Figure 4.3), resulting in a *sentiment preference matrix*.



Sentiment Score	Preference Measure
[-1, -0.05]	1
(-0.05, -0.01]	2
(-0.01, 0.01)	3
[0.01, 0.05)	4
[0.05, 1]	5

Figure 4.3 – Sentiment score distribution for all tips in English and the coarse-grained preference mapping scheme.

4.3.1.3 Coarse-grained User Preference Fusion

Having two preference matrices, the fusion criteria aim at resolving the conflict of the same entry in two matrices. The fusion framework is based on two assumptions as follows:

1. *One time check-in venues cannot reveal sufficient information about user's feeling about the venues.* In this case, sentiment preference is assumed to be more accurate and used as the final preference. For example, if a user left a very positive tip (i.e., 5 points) on a venue that she checked only once (i.e., 1 point), the final preference will be 5.
2. *A repeated customer (i.e. users who check in a venue at least twice) usually prefers the venues she visited. The preference from tips may have some impact on the overall preference.* In this case, sentiment preference is used to amend check-in preference within 1 point range as shown in Equation 4.1. More specifically, check-in preference will be increased or decreased by 1 point when sentiment preference is two points higher or less than check-in preference, respectively. For example, when a user has a preference of 3 points from check-in for a venue and left a very negative tip (1 point), the final preference will be 2 points because of the tip.

$$P_{final} = P_c + sgn(P_c - P_s) \cdot H(|P_c - P_s| - 2) \tag{4.1}$$

where P_c and P_s is the check-in and sentiment preference score, respectively. Function $sgn(x)$ is the Sign function and $H(x)$ is the Heaviside step function.

Based on the above two assumptions and the fusion criteria, we construct the hybrid preference matrix which combines both preference extracted from check-ins and tips.

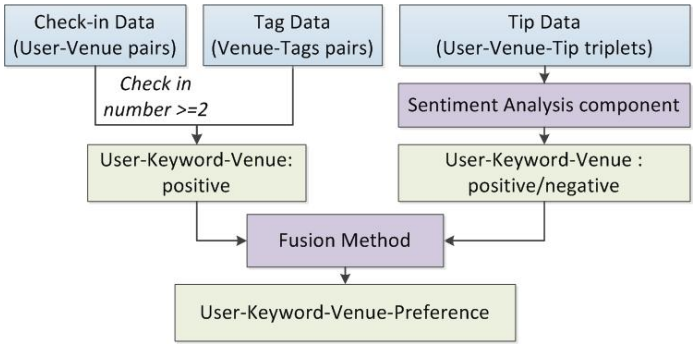


Figure 4.4 – User-Keyword-Venue tensor modeling.

Figure 4.5 – Tag cloud of keywords in "food" related venues in New York (Larger font size implies higher frequency, and vice versa.)

4.3.2 Fine-gained User Preference Modeling

In order to extract fine-grained user preference, we propose a tensor based user preference model. Figure 4.4 demonstrates the tensor modeling process. First, we extract user preference for items (in the form of keywords) on a venue (denoted as *u-k-v* preference) from two data sources, i.e., check-ins and tips, separately. Then, we propose a fusion method to merge the two sets of preference. In Foursquare, tags [2] can be used to index venues and added by users without any constraint. Thus, they might be diverse and sometimes odd, or not even a correctly spelled word. To avoid uncommon tags, we only choose the tags that have been used at least twice in all venues as keywords. The tag cloud of keywords in "food" related venues in New York is shown in Figure 4.5. Most of the popular tags that are usually used to describe restaurants (e.g., coffee, brunch, bar, burgers, pizza, etc.) are shown as expected.

2. Since tags are used in Foursquare to index venues, we do not differentiate it with the term "keywords" throughout this chapter.

4.3.2.1 User Preference from Check-in Data

A user may not necessarily like a venue if she has visited there only once, while repeated visits usually indicate she likes the place, i.e., expressing positive feedback to this venue. Based on this common sense, for each user, we select the venues being checked-in at least twice as her liked places. Since check-ins cannot provide further information on what items the users like or dislike on a venue, we assume that users have only positive feedback to all keywords at this venue. Hence, we get a set of u-k-v triplets with positive preference.

4.3.2.2 User Preference from Tip Data

Tips left by users on venues usually describe what users like or what users complain. By applying aspect-based sentiment analysis techniques, we can extract users' different opinions on different items/aspects. As shown in Figure 4.2, we also use dictionary based unsupervised sentiment analysis method. After the Noun-Phrase Chunking step, by summing up the sentiment score of each word in a phrase, we obtain the sentiment of the phrase. Here, the output of sentiment analysis on tips is a set of user-phrase-venue triplets with the corresponding sentiment score. In fine-grained user preference, we only care about the positive and negative sentiment rather than the exact sentiment scores. In order to map a phrase to keywords, we simply find out the keywords (i.e., tags) contained in a phrase. For example, the phrase "delicious pizza" is mapped to keywords "delicious" and "pizza" if they exist in the keyword set. Then, we can get a set of u-k-v triplets with positive or negative preference.

To test the accuracy of our sentiment analysis method, we randomly choose 1000 tips and manually label their fine-grain preference. The experiments give a precision of 63.91% and a recall of 88.13%. More sophisticated methods can be used to achieve better performance, but they are not our focus in this dissertation.

4.3.2.3 Fine-grained User Preference Fusion

From the check-in data, only positive preference of keywords can be extracted, while from tip data both positive and negative preference can be extracted. The user preference extracted from tips is fine-grained and contains more precise information. Hence, the fusion policy is: when the same u-k-v triplet is observed from both data sources, the preference from tips analysis is used. For example, a user checked in twice at a restaurant (tagged by burgers, pizza and beer) and left a tip complaining about the burgers there. The preference extracted from her check-in for burgers, pizza and beer in that restaurant is positive while the tip reports negative preference for burgers there. The preference extracted from tips is considered to be more accurate. Hence, the user preference for burgers in that place is

negative, and user preference for pizza and beer remains positive. Finally, to construct a user preference tensor, we assign 1 and -1 to positive and negative preference, respectively. The unknown ones are assigned 0.

4.4 Low-rank Approximation Based Personalization Algorithms

Given a coarse-grained user preference matrix and a fine-grained user preference tensor, we propose two low-rank approximation based algorithms to predict user preference in the matrix and its ranking in tensor, respectively.

4.4.1 Personalized POI Recommendation Algorithm

In this section, we present the proposed Location Based Social Matrix Factorization (LBSMF) method. First, we explain the basic principle of matrix factorization techniques, and then extend it by combining with user social network and venue similarity network for location recommendation.

4.4.1.1 Matrix Factorization

Probabilistic matrix factorization (PMF) [94] is an efficient approach in recommendation systems. It factorizes user-item rating matrix into a user-latent space matrix and an item-latent space matrix which are later used to predict the unknown ratings. Given a user-item rating matrix $R_{m\times n}$ describing m users' ratings on n items, the matrix factorization methods try to approximate $R_{m\times n}$ by a product of two matrices $U_{m\times l}$ and $V^T_{n\times l}$ which represent the user-latent space matrix and item-latent space matrix, respectively. The dimensionality of the latent space is denoted as l.

$$R_{m\times n} \approx U_{m\times l} \times V^T_{n\times l} \tag{4.2}$$

Since the rating matrix R is usually sparse in the real dataset, only the observed rating in R should be considered. In order to model the latent features of U and V, the conditional probability of the observed ratings are:

$$p(R|U,V,\sigma_R^2) = \prod_{i=1}^{m}\prod_{j=1}^{n} I_{ij}[\mathcal{N}(R_{i,j}|U_i \times V_j^T, \sigma_r^2)] \tag{4.3}$$

where I_{ij} is the indicator function that equals 1 if user i rated item j and equals 0 otherwise, $\mathcal{N}(x|\mu,\sigma^2)$ is the normal distribution with mean μ and variance σ^2. Gaussian priors are

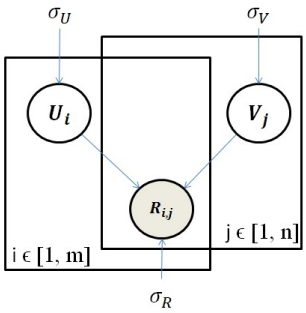


Figure 4.6 – Graphical model of probabilistic matrix factorization.

also assumed for U and V.

$$p(U|\sigma_U^2) = \prod_{i=1}^{m} [\mathcal{N}(U_i|0, \sigma_U^2 \mathbf{I})] \tag{4.4}$$

$$p(V|\sigma_V^2) = \prod_{j=1}^{n} [\mathcal{N}(V_j|0, \sigma_V^2 \mathbf{I})] \tag{4.5}$$

Based on Bayesian inference, the posterior probability of U and V are as follows.

$$p(U, V|R, \sigma_R^2, \sigma_U^2, \sigma_V^2) \propto p(R|U, V, \sigma_R^2) p(U|\sigma_U^2) p(V|\sigma_V^2) \tag{4.6}$$

By maximizing Equation 4.6, we can obtain the learned U and V for recommendation. Due to the space limit, we do not elaborate the whole derivation process and the details can be found in [94]. The graphical model of probabilistic matrix factorization is shown in Figure 4.6.

4.4.1.2 Location Based Social MF

We design our location based social MF algorithm considering both user social network and venue similarity network for location recommendation. Note that venue is considered as the item in the location recommendation. Due to social influence, we assume that a user's preference is similar to her friends', i.e. her latent features are similar to her friends'. Similarly, a venue's visiting record is similar to the similar venues (e.g., venues in the same category may probably have similar temporal traffic pattern), i.e. its latent features resemble the similar venues'. For a user i, the social influence of her friends' can be formulated as follows:

$$InfU_i = \frac{\sum_{f \in F_i} SimU_{i,f} \cdot U_f}{\sum_{f \in F_i} SimU_{i,f}} \tag{4.7}$$

where F_i is the friends set of user i and $SimU_{i,f}$ is the similarity measure between user i and her friend f. We use such similarity to determine how influential a friend is to user i. Similarly, for a venue j, the influence of the similar venues can be formulated as

$$InfV_j = \frac{\sum_{s \in N_j} SimV_{j,s} \cdot V_s}{\sum_{s \in N_j} SimV_{j,s}} \tag{4.8}$$

where N_j is the similar venues set of venue j and $SimU_{j,s}$ is the similarity measure between venue j and venue s. Note that the non-zero value in $SimU$ and $SimV$ represent the similarity measure. After normalizing each rows of $SimU$ and $SimV$ so that $\sum_{f \in F_i} SimU_{i,f} = 1$ and $\sum_{s \in N_j} SimV_{j,s} = 1$. The influence terms become:

$$\begin{aligned} InfU_i &= \sum_{f \in F_i} SimU_{i,f} \cdot U_f \\ InfV_j &= \sum_{s \in N_j} SimV_{j,s} \cdot V_s \end{aligned} \tag{4.9}$$

Based on the Gaussian priors of U and V, the latent features of users and venues are directly proportional to the combination of two factors: the zeros-means Gaussian priors as in Equation 4.4 and 4.5, and the conditional distributions of U and V given $InfU_i$ and $InfV_j$ that represent the social and inter-venue influence, which are as follows:

$$p(U|SimU, \sigma^2_{SimU}) = \prod_{i=1}^{m} [\mathcal{N}(U_i| \sum_{f \in F_i} SimU_{i,f} \cdot U_f, \sigma^2_{SimU}\mathbf{I})] \tag{4.10}$$

$$p(V|SimV, \sigma^2_{SimV}) = \prod_{j=1}^{n} [\mathcal{N}(V_j| \sum_{s \in N_j} SimV_{j,s} \cdot V_s, \sigma^2_{SimV}\mathbf{I})] \tag{4.11}$$

Such distributions ensure that a user's latent feature is close to the features of their friends, and a venue's latent feature is also close to the features of the similar venues. Hence, the conditional distribution of the latent features of U and V are:

$$\begin{aligned} &p(U|SimU, \sigma^2_U, \sigma^2_{SimU}) \propto p(U|\sigma^2_U) p(U|SimU, \sigma^2_{SimU}) \\ &= \prod_{i=1}^{m} [\mathcal{N}(U_i|0, \sigma^2_U\mathbf{I})] \times \prod_{i=1}^{m} [\mathcal{N}(U_i| \sum_{f \in F_i} SimU_{i,f} \cdot U_f, \sigma^2_{SimU}\mathbf{I})] \end{aligned} \tag{4.12}$$

$$\begin{aligned} &p(V|SimV, \sigma^2_V, \sigma^2_{SimV}) \propto p(V|\sigma^2_V) p(V|SimV, \sigma^2_{SimV}) \\ &= \prod_{j=1}^{n} [\mathcal{N}(V_j|0, \sigma^2_V\mathbf{I})] \times \prod_{j=1}^{n} [\mathcal{N}(V_j| \sum_{s \in N_j} SimV_{j,s} \cdot V_s, \sigma^2_{SimV}\mathbf{I})] \end{aligned} \tag{4.13}$$

Similar to Equation 4.6, using Bayesian inference the posterior probability of latent features is:

$$
\begin{aligned}
& p(U,V|R,SimU,SimV,\sigma_R^2,\sigma_U^2,\sigma_V^2,\sigma_{SimU}^2,\sigma_{SimV}^2) \\
& \propto p(R|U,V,\sigma_R^2)p(U|SimU,\sigma_U^2,\sigma_{SimU}^2)p(V|SimV,\sigma_V^2,\sigma_{SimV}^2) \\
& = \prod_{i=1}^{m}\prod_{j=1}^{n} I_{ij}[\mathcal{N}(R_{i,j}|g(U_i \times V_j^T),\sigma_r^2)] \\
& \times \prod_{i=1}^{m}[\mathcal{N}(U_i|\sum_{f\in F_i} SimU_{i,f}\cdot U_f,\sigma_{SimU}^2\mathbf{I})] \\
& \times \prod_{j=1}^{n}[\mathcal{N}(V_j|\sum_{s\in N_j} SimV_{j,s}\cdot V_s,\sigma_{SimV}^2\mathbf{I})] \\
& \times \prod_{i=1}^{m}[\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})] \times \prod_{j=1}^{n}[\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})]
\end{aligned}
\tag{4.14}
$$

where $g(x)$ is the logistic function that bounds the range of predictions into [0, 1]. In order to keep the generality, the user-venue ratings are mapped to interval [0, 1] using the function $f(x) = (x-1)/(max\_rating-1)$, and recovered later using $f^{-1}(x)$. Then, the log posterior probability of Equation 4.14 is:

$$
\begin{aligned}
& \ln p(U,V|R,SimU,SimV,\sigma_R^2,\sigma_U^2,\sigma_V^2,\sigma_{SimU}^2,\sigma_{SimV}^2) \\
& = -\frac{1}{2\sigma_R^2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}[R_{i,j} - g(U_i \times V_j^T)] \\
& - \frac{1}{2\sigma_{SimU}^2}\sum_{i=1}^{m}(U_i - \sum_{f\in F_i} SimU_{i,f}U_f)(U_i - \sum_{f\in F_i} SimU_{i,f}U_f)^T \\
& - \frac{1}{2\sigma_{SimV}^2}\sum_{j=1}^{n}(V_j - \sum_{s\in N_j} SimV_{j,s}V_s)(V_j - \sum_{s\in N_j} SimV_{j,s}V_s)^T \\
& - \frac{1}{2}[\frac{1}{\sigma_U^2}\sum_{i=1}^{m} U_iU_i^T + \frac{1}{\sigma_V^2}\sum_{j=1}^{n} V_jV_j^T + (\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij})\ln\sigma_R^2] \\
& - \frac{1}{2}[ml(\ln\sigma_U^2 + \ln\sigma_{SimU}^2) + nl(\ln\sigma_V^2 + \ln\sigma_{SimV}^2)] + \mathcal{C}
\end{aligned}
\tag{4.15}
$$

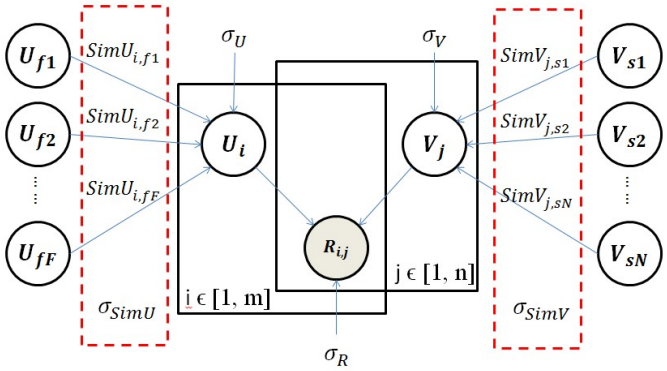


Figure 4.7 – Graphical model of LBSMF.

We aim at maximizing log posterior probability of U and V keeping the variance parameter fixed. Maximizing above term is equivalent to minimizing the following objective function:

$$\begin{aligned}
&\mathcal{L}(R, SimU, SimV, U, V) \\
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}[R_{i,j} - g(U_i \times V_j^T)] + \frac{1}{2}[\lambda_U \sum_{i=1}^{m} U_i U_i^T + \lambda_V \sum_{j=1}^{n} V_j V_j^T] \\
&+ \frac{1}{2}\alpha \sum_{i=1}^{m}(U_i - \sum_{f\in F_i} SimU_{i,f}U_f)(U_i - \sum_{f\in F_i} SimU_{i,f}U_f)^T \\
&+ \frac{1}{2}\beta \sum_{j=1}^{n}(V_j - \sum_{s\in N_j} SimV_{j,s}V_s)(V_j - \sum_{s\in N_j} SimV_{j,s}V_s)^T
\end{aligned} \tag{4.16}$$

where $\lambda_U = \sigma_R^2/\sigma_U^2$, $\lambda_V = \sigma_R^2/\sigma_V^2$, $\alpha = \sigma_R^2/\sigma_{SimU}^2$ and $\beta = \sigma_R^2/\sigma_{SimV}^2$. Applying the gradient descent approach on each user latent feature vector U_i and venue latent feature vector V_j for above objective function, we have

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial U_i} &= \sum_{j=1}^{n} I_{ij} V_j g'(U_i \times V_j^T)[g(U_i \times V_j^T) - R_{i,j}] + \lambda_U U_i + \alpha(U_i - \sum_{f\in F_i} SimU_{i,f}U_f) \\
&- \alpha \sum_{\{f|i\in F_f\}} simU_{f,i}(U_f - \sum_{w\in F_f} SimU_{f,w}U_w)
\end{aligned} \tag{4.17}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial V_j} &= \sum_{i=1}^{m} I_{ij} U_i g'(U_i \times V_j^T)[g(U_i \times V_j^T) - R_{i,j}] + \lambda_V V_j + \beta(V_j - \sum_{s\in N_j} SimV_{j,s}V_s) \\
&- \beta \sum_{\{s|j\in N_s\}} simV_{s,j}(V_s - \sum_{p\in N_s} SimV_{s,p}V_s)
\end{aligned} \tag{4.18}$$

where $g'(x) = e^{-x}/(1+e^{-x})^2$ which is the derivative of the logistic function. Using gradient descent approach, U_i and V_j are updated in each iteration according to Equation

4.17 and 4.18, respectively. The graphical model of our proposed location based social matrix factorization method is illustrated in Figure 4.7. Compared to the PMF model, we introduce the user friendship network and venue similarity network in matrix factorization approach in order to consider the influence of inter-user and inter-venue relationships in location recommendation.

4.4.2 Personalized POI Search Algorithm

In this section, we present the Multi-Tuple based Ranking Tensor Factorization (MT-RTF) algorithm. The purpose of MT-RTF algorithm is to rank venues in the order of user preferred, with unknown preference, and with negative preference. To achieve this goal, MT-RTF algorithm predicts the ranking of user preference for venues in tensor. First, we select an appropriate tensor factorization model. Then, based on this model we define an objective function which measures the multi-tuple ranking quality. Finally, we extend the learning framework in [116] to maximize the objective function in the learning process.

4.4.2.1 Tensor Factorization Model

Tensor factorization techniques intend to decompose a tensor into multiple factors. For the *u-k-v* tensor, let $\hat{U}$, $\hat{K}$ and $\hat{V}$ denote the user, keyword and venue feature matrices, with dimension of $|U| * l$, $|K| * l$ and $|V| * l$, respectively. Note that l is called latent space dimension (or factorization dimension) which is the most important parameter in tensor factorization. It controls the number of features used in the factorization process. The U, K, V are finite sets of users, keywords and venues, respectively. The decomposition can be formulated as:

$$\hat{Y} = \hat{C} \times_U \hat{U} \times_K \hat{K} \times_V \hat{V} \tag{4.19}$$

where $\times_n$ is the mode-n tensor product with matrix. The core tensor $\hat{C}$ with dimension $l * l * l$ handles the correlation among different factors. The value of each element in $\hat{Y}$ is calculated as:

$$\hat{y}_{u,k,v} = \sum_{\hat{u}} \sum_{\hat{k}} \sum_{\hat{v}} \hat{c}_{\tilde{u},\tilde{k},\tilde{v}} \cdot \hat{u}_{u,\tilde{u}} \cdot \hat{k}_{k,\tilde{k}} \cdot \hat{v}_{v,\tilde{v}} \tag{4.20}$$

where $\tilde{u}, \tilde{k}, \tilde{v} \in \{1, ..., l\}$ are indices of latent space. This model is called Tucker decomposition model [137]. If we set the core tensor as a diagonal tensor:

$$\hat{c}_{\tilde{u},\tilde{k},\tilde{v}} = \begin{cases} 1, \ if \quad \tilde{u} = \tilde{k} = \tilde{v} \\ 0, \ else \end{cases} \tag{4.21}$$

We obtain a Canonical decomposition model with each element calculated as:

$$\hat{y}_{u,k,v} = \sum_{\tilde{f}} \hat{u}_{u,\tilde{f}} \cdot \hat{k}_{k,\tilde{f}} \cdot \hat{v}_{v,\tilde{f}} \tag{4.22}$$

where $\tilde{f} \in \{1, ..., l\}$ is the indices of latent space. As a special case of Canonical decomposition model, the pairwise interaction model [117] explicitly captures the pairwise interaction among the three factors:

$$\hat{y}_{u,k,v} = \sum_{\tilde{f}} \hat{u}^K_{u,\tilde{f}} \cdot \hat{k}^U_{k,\tilde{f}} + \hat{u}^V_{u,\tilde{f}} \cdot \hat{v}^U_{v,\tilde{f}} + \hat{k}^V_{k,\tilde{f}} \cdot \hat{v}^K_{v,\tilde{f}} \tag{4.23}$$

where $\hat{u}^K$ represent the interaction between user and keyword from user's perspective, and so on. When predicting venue ranking, the interaction between user and keyword vanishes. Using vector representation we get:

$$\hat{y}_{u,k,v} = \hat{u}_u \cdot (\hat{v}^U_v)^T + \hat{k}_k \cdot (\hat{v}^K_v)^T \tag{4.24}$$

where $\hat{u}_u$ and $\hat{k}_k$ are the feature vectors in $\hat{U}$ and $\hat{K}$. Additionally, $\hat{v}^U_v$ and $\hat{v}^K_v$ are the feature vectors in $\hat{V}^U$ and $\hat{V}^K$. Note that in this model, a tensor is decomposed into four factors, i.e., $\hat{U}$, $\hat{K}$, $\hat{V}^U$ and $\hat{V}^K$. This model is used in our work to factorize the *u-k-v* tensor.

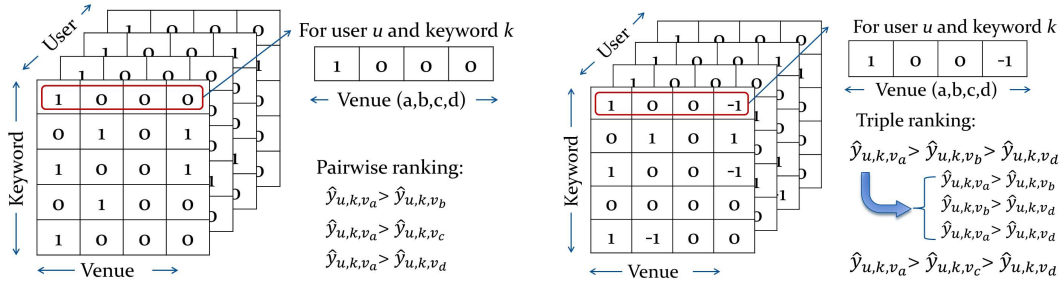
4.4.2.2 Optimization criterion

Optimization criterion is represented by an objective function when performing tensor factorization. Existing approaches can only handle positive preference [115, 117]. Given a user u and a keyword k, those works aim at ranking user preferred venues in front of others, which can be formulated as:

$$Obj_{po} = \sum_{v^+ \in V^+_{u,k}} \sum_{v^0 \in V^0_{u,k}} (\hat{y}_{u,k,v^+} - \hat{y}_{u,k,v^0}) \tag{4.25}$$

where $V^+_{u,k}$ and $V^0_{u,k}$ represent venues with positive preference and with unknown preference, respectively. Maximizing the above function is able to rank the venues with positive preference in front of the others, as shown in Figure 4.8(a), where rank of venue a is higher than that of venue b, c and d. However, since our *u-k-v* tensor further includes negative preference, this objective function cannot handle such case. As shown in Figure 4.8(b), for a user u and a keyword k, the rank of venue a (with positive preference) is higher than that of venue b and c (unknown preference), and the rank of venue b and c are higher than that of venue d (with negative preference). Then, each triple ranking relation can be seen as three pairwise ranking relations, as shown in Figure 4.8(b). Let $V^-_{u,k}$ denote venues with negative preference. In addition to Equation 4.25, the two other pairwise ranking relations are between $V^0_{u,k}$ and $V^-_{u,k}$, $V^+_{u,k}$ and $V^-_{u,k}$.

$$Obj_{on} = \sum_{v^0 \in V^0_{u,k}} \sum_{v^- \in V^-_{u,k}} (\hat{y}_{u,k,v^0} - \hat{y}_{u,k,v^-}) \tag{4.26}$$



(a) Pairwise ranking relation

(b) Triple ranking relation

Figure 4.8 – Multi-Tuple Ranking Scheme.

$$Obj_{pn} = \sum_{v^+ \in V^+_{u,k}} \sum_{v^- \in V^-_{u,k}} (\hat{y}_{u,k,v^+} - \hat{y}_{u,k,v^-}) \tag{4.27}$$

The optimization is then performed successively for the three pairwise ranking relations. Let $\vec{v}_{u,k}$ denote the venue vector given the user u and keyword k. The vector $\vec{v}_{u,k}$ belongs to *pairwise ranking* if $\vec{v}_{u,k}$ only contains two different values, while $\vec{v}_{u,k}$ belongs to *triple ranking* if $\vec{v}_{u,k}$ contains three different values, i.e., 1, -1 and 0. Considering both pairwise and triple ranking relations, the optimization criterion of MT-RTF is defined as:

$$Obj = \begin{cases} Obj_{xx}, \ if \ \ \vec{v}_{u,k} \in pairwise \ ranking \\ Obj_{po} + Obj_{on} + Obj_{pn}, \ if \ \ \vec{v}_{u,k} \in triple \ ranking \end{cases} \tag{4.28}$$

where Obj_{xx} represents Obj_{po}, Obj_{on} or Obj_{pn} for $\vec{v}_{u,k}$ containing 1/0, 0/-1 or 1/-1, respectively. By maximizing Obj for all observed (u, k) pairs, we get finally the objective function as:

$$\max_{\hat{U},\hat{K},\hat{V}^U,\hat{V}^K} \sum_{\{(u,k)|\exists v, y_{u,k,v} \neq 0\}} Obj \tag{4.29}$$

4.4.2.3 Learning Process

We adopt Bayesian personalized ranking learning algorithm [116] as the learning framework. A bootstrap sampling method is used to reduce learning time. Since our objective function considers all the data in tensor, to target the data obtained from sampling approach, we extract an atomic objective function for each ranking venue pair, denoted as $\hat{y}_{u,k,v^a,v^b}$ for a given ranking pair $\hat{y}_{u,k,v^a}$ and $\hat{y}_{u,k,v^b}$, where $(a, b) \in \{(+, 0), (0, -), (+, -)\}$.

$$\hat{y}_{u,k,v^a,v^b} = (\hat{y}_{u,k,v^a} - \hat{y}_{u,k,v^b}) \tag{4.30}$$

Algorithm 4.1 MT-RTF Learning Algorithm

Input: T, Θ

1: initialize Θ
2: **repeat**
3: draw a (u, k) pair uniformly from T
4: **if** $\vec{v}_{u,k} \in$ pairwise ranking **then**
5: draw (v^a, v^b) uniformly from $\vec{v}_{u,k}$
6: $\rho = (1 - g'(\hat{y}_{u,k,v^a,v^b}))$
7: $\Theta = \Theta + \alpha \cdot [\rho \cdot \frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \Theta} - \lambda \cdot \Theta]$
8: **end if**
9: **if** $\vec{v}_{u,k} \in$ triple ranking **then**
10: draw (v^+, v^0, v^-) uniformly from $\vec{v}_{u,k}$
11: **for** $(v^a, v^b) \in \{(v^+, v^0), (v^0, v^-), (v^+, v^-)\}$ **do**
12: $\rho = (1 - g'(\hat{y}_{u,k,v^a,v^b}))$
13: $\Theta = \Theta + \alpha \cdot [\rho \cdot \frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \Theta} - \lambda \cdot \Theta]$
14: **end for**
15: **end if**
16: **until** convergence of Obj
17: **return** $\hat{\Theta}$

Gradient descent approach is used to update parameters $\hat{U}, \hat{K}, \hat{V}^U$ and $\hat{V}^K$ in each iteration. Combining with Equation 4.24, the gradients of $\hat{y}_{u,k,v^a,v^b}$ are:

$$\frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \hat{u}_u} = (\hat{v}^U_{v^a} - \hat{v}^U_{v^b}), \frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \hat{k}_k} = (\hat{v}^K_{v^a} - \hat{v}^K_{v^b}) \tag{4.31}$$

$$\frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \hat{v}^U_{v^a}} = \hat{u}_u, \frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \hat{v}^U_{v^b}} = -\hat{u}_u \tag{4.32}$$

$$\frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \hat{v}^K_{v^a}} = \hat{k}_k, \frac{\partial \hat{y}_{u,k,v^a,v^b}}{\partial \hat{v}^K_{v^b}} = -\hat{k}_k \tag{4.33}$$

Given a tensor T and a set of parameters Θ i.e., $\hat{U}, \hat{K}, \hat{V}^U$ and $\hat{V}^K$, MT-RTF learning algorithm is illustrated in Algorithm 4.1. Note that $g(x) = \frac{1}{1+e^{-x}}$ is the logistic function. The α controls the learning step and λ is the regularization parameter. In each iteration, we first select one (u, k) pair randomly (Line 3), and then randomly select pairwise ranking relation (Line 4-5) or triple ranking relation (Line 9-10) according to $\vec{v}_{u,k}$. For pairwise ranking, the optimization is conducted only for v^a, v^b (Line 6-7). For triple ranking, the optimization is conducted successively for $\{(v^+, v^0), (v^0, v^-), (v^+, v^-)\}$ (Line 11-14). The algorithm converges until no further improvement for the objective function Obj. The output of MT-RTF is the optimized Θ. Using Equation 4.24, the predicted ranking score

Table 4.1 – Dataset statistics for personalized POI recommendation.

Dataset	New York Restaurant	London
Users	2601	1233
Venues	2392	1623
Density using check-in	0.0042	0.0048
Density using HPM	0.0053	0.0058
Social network density	0.0007	0.0029

can be obtained. Based on such ranking score, for a given user and a keyword, venues can be ranked.

4.5 Experimental Evaluation

Using the real world user activity data collected from LBSNs, we experimentally evaluate the proposed framework. Specifically, we focus on two types of user preference (coarse-grained and fine-grained user preference), which are then applied in different application scenarios (i.e., personalized POI recommendation and search), respectively.

4.5.1 Dataset Description

In this work, we use a collection of Foursquare check-ins lasting for 4 months (from 24 October 2011 to 20 February 2012). Since we only process tips written in English for sentiment analysis, we select specific urban scale datasets in English-speaking countries for evaluation.

4.5.1.1 Coarse-grained User Preference Matrix

In order to build coarse-grained user preference matrices, we select user activity data in two cities, i.e. New York and London. We choose the food related venue check-ins ("Food" root category, containing 86 sub-categories such as French restaurant, Italian restaurant, etc.) in New York (denoted as New York Restaurant) and keep all categories in London. Moreover, in Foursquare, user relationship is not public available. We indirectly build social network via twitter follower and following relationship, i.e. we assume that the friendship exists if two users follow each other in Twitter. The data statistics is shown in Table 4.1.

4.5.1.2 Fine-grained User Preference Tensor

In order to build a fine-grained user preference tensor, we select the New York Restaurant dataset. We do not use the London dataset to build its fine-grained user preference tensor, due to fact that there is very limited tag data in London dataset and tag data is

Table 4.2 – Dataset statistic for personalized POI Search.

User number	994
Keyword number	728
Venue number	1008
Number of the observed u-k-v triplets	51091
Data density	0.007%
Positive feedback number	43924
Negative feedback number	7167

necessary for check-in user preference augmentation. To get a relatively dense tensor in experiments, we select 20-core data [3], resulting in a *u-k-v* tensor with dimensionality of (994*728*1008). The data statistics is shown in Table 4.2.

4.5.2 POI Recommendation with Coarse-grained User Preference

In order to validate the proposed POI recommendation approach, we evaluate the proposed preference model and algorithm for POI recommendation, and compare it with other state-of-the-art methods. Our evaluation tries to address the following questions:

1. How does the proposed hybrid preference model capture users' preference? Can it maintain the consistency of the preference extracted from check-ins and tips?
2. Comparing with other methods, does LBSMF achieve better performance?
3. How do social network and venue similarity network affect recommendation performance and to what extent?

4.5.2.1 Social and Inter-venue Influence Modeling

As inputs to LBSMF, social network and venue similarity network need to be built properly. As mentioned previously, social network is extracted based on user follower/following relationship. Since we have the preference of all the users, the evaluation of similarity between two users can then be calculated by measuring the preference vectors of these two users. Similar to [87], Pearson Correlation Coefficient is used as similarity measure in this study.

With regard to venue similarity network, we extract venue category information from Foursquare to build a 0/1 based venue similarity network. For two venues, the similarity score is set to 1 if both venues have the same sub-category in Foursquare, it is set to 0 if there is no overlapping sub-category. Since our experiment dataset is constrained to these

3. The p-core of a tensor is the largest subset of the tensor with the property that every user, every keyword and every venue has to occur in at least p records.

two cities, the geographical influence is omitted in this experiment. It will be considered in the future work combining with venue semantic similarity from tips.

4.5.2.2 Evaluation Metrics

Two common metrics are used for evaluation: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

$$MAE = \frac{1}{|T|} \sum_{R_{i,j} \in T} |R_{i,j} - \hat{R}_{i,j}| \tag{4.34}$$

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{R_{i,j} \in T} (R_{i,j} - \hat{R}_{i,j})^2} \tag{4.35}$$

where T is the test dataset. $R_{i,j}$ and $\hat{R}_{i,j}$ represent the observed preference and the predicted preference measure of user i on venue j, respectively. Smaller MAE and RMSE imply better performance. The greater difference between them, the greater the variance in the individual errors in the test set.

4.5.2.3 Hybrid Preference Model Evaluation

In order to evaluate the proposed hybrid preference model using both check-ins and tips, we compare the performance of LBSMF with different preference models. A model built from only user's check-in behavior is used as baseline. Obviously, considering tip data can increase the density of the preference matrix. In order to prove that the hybrid preference model outperforms other models not merely because it alleviates the sparsity problem, we proposed a null model with the same density and same distribution of the preference record numbers. Hence, the models used and tested in this experiment are as follows:

— Basic model (BM) that only uses *check-in preference matrix*.

— Tip null model (TNM) that considers tips influence in a random way. It shuffles randomly the preference measure in *sentiment preference matrix* and then fuses it with *check-in preference matrix*. In this way, it preserves the same distribution of the number of preference records.

— Hybrid preference model (HPM) that uses our proposed *hybrid preference matrix*.

We fixed $\lambda_U = \lambda_V = 0.005$, *learning rate* $= 0.02$ for all the evaluations conducted in the following section. The social and inter-venue influence parameters are set as $\alpha = 0.001$ and $\beta = 0.01$ for New York Restaurant dataset, $\alpha = 0.002$ and $\beta = 0.02$ for London dataset because they result in the best performance (the detailed study about parameter tuning is

Table 4.3 – Comparison between different preference models.

Dataset	Training	Metric	BM	TNM	HPM
New York Restaurant	90%	RMSE	1.0137	0.8887	0.8524
		MAE	0.8072	0.7032	0.6204
	80%	RMSE	1.0386	**1.0506**	0.9580
		MAE	0.8103	**0.8306**	0.7345
London	90%	RMSE	1.1045	0.9864	0.8929
		MAE	0.9031	0.7889	0.7022
	80%	RMSE	1.1245	1.0895	1.0119
		MAE	0.9147	0.8828	0.8075

presented in evaluation of social and inter-venue influence). We use different percentage of data (i.e. 90%, 80%) for training. For example, training data 90% means that we randomly select 90% of the preference records as the training set, and the rest 10% as the test set. The latent space dimension is set to 10 in this experiments. The results are shown in Table 4.5.2.3. Each result is the mean value of five repeated trials.

We can observe clearly that HPM achieves the best performance for both dataset. The BM which only considers check-in data yields the worst performance among the three models. Although TNM model has the same density and the same distribution of the preference record numbers as HPM, the performance is still poorer than HPM. An interesting observation is that TNM model is even worse than BM when using New York Restaurant dataset with 80% of data as training set. We can see even if TNM increases the density of the preference matrix but it impacts dramatically on user's real preference due to the random assignment of sentiment preference measure.

These observations strongly support that the proposed HPM is able to characterize users' preference and maintain the consistency of user preference modeled from both check-in and tip data.

4.5.2.4 Location Recommendation Evaluation

In this section, we compare our proposed LBSMF with the following approaches to show its effectiveness in location recommendation.

— Classical Collaborative filtering (CF) is used as baseline.

— Probabilistic matrix factorization (PMF) [94]: one classical matrix factorization approach. Our approach extends PMF by introducing social and inter-venue influence.

— SocialMF [61]: this approach considers social network influence in recommendation problem and treats friend's impact equally. After a series of experiments, the social influence parameter is set to 0.01 since it achieves best results on both of our datasets.

— Social Regularized MF (SRMF) [87]: it considers not only social network connection,

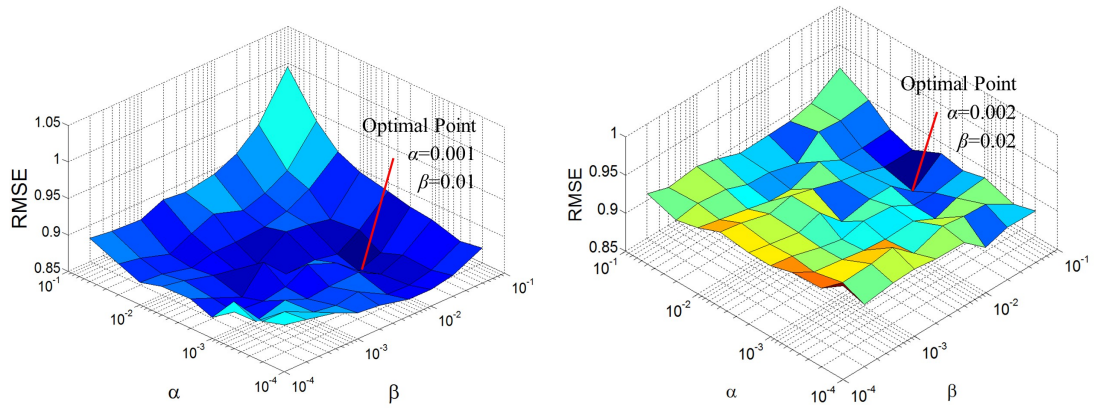
Table 4.4 – Performance comparisons with other approaches.

Dataset	Training	Metric	Dimension = 5					Dimension = 10				
			CF	PMF	SocialMF	SRMF	LBSMF	CF	PMF	SocialMF	SRMF	LBSMF
New York Restaurant	90%	RMSE	1.2463	0.9440	0.9364	0.9342	0.9184	1.2463	0.9136	0.8889	0.8755	0.8524
		Improve	26.31%	2.71%	1.92%	1.69%		31.61%	6.70%	4.11%	2.64%	
		MAE	0.7190	0.7182	0.7074	0.7034	0.6949	0.7190	0.7047	0.6429	0.6238	0.6204
		Improve	3.35%	3.24%	1.77%	1.21%		13.71%	11.96%	3.50%	0.55%	
	80%	RMSE	1.4887	1.0209	1.0279	1.0206	1.0040	1.4887	0.9942	0.9748	0.9713	0.9580
		Improve	32.56%	1.66%	2.33%	1.63%		35.65%	3.64%	1.72%	1.37%	
		MAE	0.8435	0.8262	0.8204	0.7959	0.7916	0.8435	0.8101	0.7585	0.7425	0.7345
		Improve	6.15%	4.19%	3.51%	0.54%		12.92%	9.33%	3.16%	1.08%	
London	90%	RMSE	1.3787	0.9758	0.9651	0.9519	0.9328	1.3787	0.9763	0.9125	0.9382	0.8929
		Improve	32.34%	4.41%	3.35%	2.01%		35.24%	8.54%	2.15%	4.83%	
		MAE	0.8687	0.7719	0.7682	0.7568	0.7315	0.8687	0.7882	0.7203	0.7379	0.7022
		Improve	15.79%	5.23%	4.78%	3.34%		19.17%	10.91%	2.51%	4.84%	
	80%	RMSE	1.6222	1.0733	1.0497	1.0547	1.0273	1.6222	1.0496	1.0358	1.0440	1.0119
		Improve	36.67%	4.29%	2.13%	2.60%		37.62%	3.59%	2.31%	3.07%	
		MAE	1.0441	0.8682	0.8539	0.8520	0.8266	1.0441	0.8508	0.8246	0.8441	0.8075
		Improve	20.83%	4.79%	3.20%	2.98%		22.66%	5.09%	2.07%	4.34%	

but also the similarity measure between friends. We implement the individual-based regularization model using Pearson Correlation Coefficient as similarity measure in the experiment since it reports the best results. Note that the social regularization term is added in a different way from that of SocialMF. Similar to SocialMF, the social influence parameter is set to 10^{-6} for the best performance.

The dimension of latent space is set to 5 and 10, respectively. Other parameters are set as the same as in the previous section. The results are reported in Table 4.4. Each result is the average value of five repeated experiments. No matter 5-dimension or 10-dimension representation of latent space is used, the gain of LBSMF is significant comparing to other approaches. Considering inter-venue influence, both datasets achieve better RMSE and MAE. Besides the RMSE and MAE value, the rate of improvement over other approaches is also indicated in Table 4.4.

As can be seen from Table 4.4, the traditional CF performs the worst. The PMF method achieves better results comparing to CF. Considering social influence, both SocialMF and Social Regularized MF perform better than those methods that ignore social influence, which confirms that social influence is able to impact user preference behavior to some extent. LBSMF that takes both social and inter-venue relationship into account achieves the best results comparing to the state-of-the-art approaches. The results also imply that inter-venue influence such as category in this experiment has strong influence on location recommendation.



(a) New York restaurant dataset.

(b) London dataset.

Figure 4.9 – Impact of parameters α and β (Dimension=10, Training data 90%).

4.5.2.5 Social and Inter-venue Influence

LBSMF approach leverages the parameters α and β to control the influence from social network and venue similarity network, respectively. In this section, we investigate the impact of parameters α and β. Keeping latent space dimension as 10, training data 90%, we set parameters α and β varying within [0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05], and use RMSE as metric. Smaller value of α or β implies that we consider less social or inter-venue influence, and vice versa. In the extreme case that we set α and β to zero, LBSMF approach becomes PMF because it only uses users' preference for recommendation. On the contrary, a large value for α or β implies that social network or venue similarity dominates the latent feature learning process.

Figure 4.9 plots the RMSE metric under different α and β setting for both New York Restaurant and London datasets. Obviously, there is a concave surface of RMSE values for each dataset. Take the evaluation results with New York restaurant dataset as example, considering the most social influence and the least inter-venue influence corresponds to the left corner ($\alpha = 0.05$ and $\beta = 10^{-4}$) of the Figure 4.9(a), which has a relatively high RMSE measure. Similar situation is observed for the right corner ($\alpha = 10^{-4}$ and $\beta = 0.05$) when considering the most inter-venue influence and the least social influence. Moreover, if the recommendation is mainly based on social and inter-venue influence while considering the least user's own preference, the result becomes the worst ($\alpha = 0.05$ and $\beta = 0.05$). On the other hand, when considering little social and venue impact, the RMSE achieves almost the same result as PMF ($\alpha = 10^{-4}$ and $\beta = 10^{-4}$).

The optimal point can then be found when the lowest RMSE value achieved. For New York restaurant dataset, the optimal point (RMSE = 0.8524) is achieved at $\alpha = 0.001$ and

$\beta = 0.01$. For London dataset, it achieved at $\alpha = 0.002$ and $\beta = 0.02$ (RMSE = 0.8929).

4.5.3 POI Search with Fine-grained User Preference

For personalized search, evaluation is not an easy task because the returned results can be judged only by the searchers themselves. Obviously, such an approach is costly in our case because it is difficult to interview Foursquare users by questionnaire. Therefore, we evaluate the personalized ranking quality, i.e., whether the top ranked results contain more user liked venues and less disliked venues. The performance evaluation intends to answer the following questions:

1. How does the latent space dimension influence venue ranking performance?
2. Can our approach achieve better performance compared with the state-of-the-art approaches? What advantages can be brought out by considering fine-grained and negative preference?
3. Does our approach perform consistently for different types of users?

4.5.3.1 Evaluation Plan and Metric

To answer these questions, we first test MT-RTF performance using different latent space dimensions. By fixing to one latent space dimension, we then compare its performance with other state-of-the-art approaches. Finally, we show the performance of MT-RTF algorithm for different types of users.

A test set S is constructed by randomly selecting u-k pairs and all related venues, i.e., $\vec{v}_{u,k}$. The remaining is used as the training set. ($\vec{v}_{u,k}$ is set to 0 in training set for those u-k pairs selected by the test set). Using MT-RTF on the training set to perform the venue ranking, the predicted ranking for u-k pairs in the test set S is then evaluated.

The classical evaluation metrics in IR (Information Retrieval) often evaluate whether a result is relevant or not. However, in our case, for a given u-k pair, the venues may fall into three categories, i.e., venues with positive, negative, or unknown preference. While the ones with positive or unknown preference can be treated as "relevant" or "non-relevant", the negative ones cannot be simply considered as "non-relevant". Because putting a user disliked venue on the top will decrease user experience more than a non-relevant venue. Hence, by adjusting Mean Average Precision (MAP) which is a widely used metric in IR community, we introduce a metric named *Mean Average Satisfaction* (MAS). To introduce MAS, we first explain the definition of MAP. For a test set S, MAP is defined as follows:

$$MAP = \sum_{(u,k)\in S} \frac{\sum_{i=1}^{n} \sum_{j=1}^{i} \frac{r(j)}{i} \cdot r(i)}{N^{+}} \tag{4.36}$$

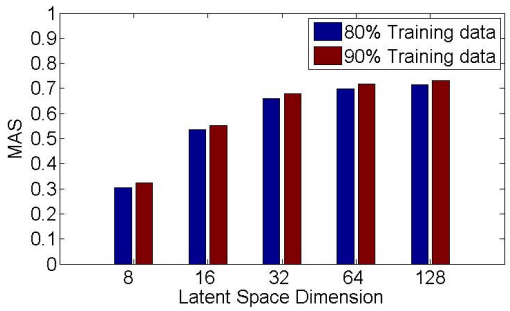


Figure 4.10 – Performance with different latent space dimensions.

where N^+ and n are the number of relevant venues (i.e., venues with positive preference) and number of retrieved venues, respectively. The relevance function $r(i)$ is set to 1 if the i^{th} venue in the results is relevant and 0 otherwise. Since this definition does not consider the venues with negative preference, the extension of MAS comparing with MAP is to introduce a punishment against ranking user disliked venues on the top. Its definition is as follows:

$$MAS = \sum_{(u,k)\in S} \frac{\sum_{i=1}^{n} \sum_{j=1}^{i} \frac{sat(j)}{i} \cdot r(i)}{N^+} \tag{4.37}$$

where the satisfaction function $sat(i)$ is set to 1, 0 or -1 if the i^{th} venue is the one with positive, unknown or negative preference, respectively. A higher value implies the top results contain more venues with positive preference and fewer venues with negative preference. Hence, MAS can be regarded as an indicator of user experience for the retrieved venue ranking.

4.5.3.2 Performance Test with Different Latent Space Dimensions

In this experiment, we choose 80% and 90% of the dataset as training set, and then vary the latent space dimension in the order of 8, 16, 32, 64 and 128. We empirically set the learning step α to 0.1 and regularization parameter λ to 0.00001. In all the performance tests, each result is the mean value of ten repeated trials. Figure 4.10 reports the results. With the increase of the latent space dimension, the ranking performance of MT-RTF increases. A slight improvement is observed for using 90% of the data as training set comparing the case of using 80%. We also find that no significant improvement of MAS for dimension higher than 64, which indicates the convergence of the algorithm in terms of latent space dimension. Hence, in the following experiments, the latent space dimension is fixed as 64 and training data percentage is set to 90%.

4.5.3.3 Comparison with Other Approaches

In order to further validate the effectiveness of MT-RTF, we compare it with the existing personalized search approaches shown below:

— PopularK: for a given keyword, venues are ranked by its popularity in a descending order, regardless of users. This is deemed as a non-personalized search approach because it returns the identical search results to all users.

— Relevance+PrefU: for a given keyword and a user, venues are firstly filtered by venue-keyword relevance and then re-ranked by the user preference on venues and keywords. This can be regarded as a typical personalized search approach using coarse-grained user preference.

— HOSVD: high order singular value decomposition [42] which performs the low-rank approximation. It corresponds to a Tucker decomposition optimized for square-loss.

— PITF: pairwise interaction tensor factorization [117] which only incorporates positive preference into factorization. Using this approach, we consider negative preference (i.e., -1) as unknown preference (i.e., 0) in the training set in order to ignore negative preference.

We set latent space dimension as 64 for all the tensor based approaches, i.e., MT-RTF, HOSVD and PITF, and keep other parameters the same as in the previous section. Firstly, we report the overall performance on the whole test set (denoted as T_ALL). In order to deeply investigate the improvement of considering negative preference, we choose the partial test set that only contains u-k pairs with negative preference (T_NEG) to show what performance can be achieved.

The left part of Figure 4.11 illustrates the overall performance. Obviously, all personalized approaches outperform the non-personalized search PopularK, which indicates that personalization is able to enhance user experience, i.e., leading higher MAS. Among the personalized approaches, HOSVD performs the worst. This might be caused by two reasons, viz. the tensor sparsity problem and weakness of HOSVD for ranking problem. Relevance+PrefU approach that considers coarse-grained user preference of venues performs better than HOSVD but still gets unsatisfactory results. The high performance of MT-RTF and PITF proves that the ranking tensor factorization approach is efficient in solving such ranking problem. Furthermore, MT-RTF that considers both positive and negative preference achieves higher performance comparing with PITF.

The right part of Figure 4.11 illustrates the performance for T_NEG. The proposed MT-RTF outperforms other approaches. Considering negative preference can significantly improve the user experience for those users with negative preference. In our dataset, the total number of observed negative preference (7167) is only 1/6 of the positive ones (43924).

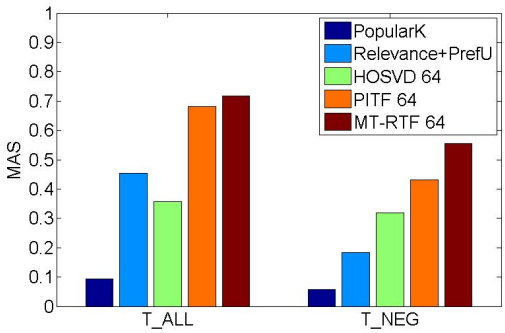


Figure 4.11 – Performance comparison with other approaches.

Such statistic explains that the improvement of MT-RTF for all users is not as much as that for users with negative preference. However, we believe that with more data collected in the future, the number of users with negative preference will increase. Thus, the advantages of MT-RTF will become more significant. An interesting observation is that the performance of Relevance+PrefU dramatically decreases when tackling negative preference. Because *coarse-grained user preference* on venue fails in the case that user has both positive and negative preference in one venue. On the contrary, *fine-grained user preference* can fully capture such detailed preference.

4.5.3.4 Performance Test for Different Types of Users

In LBSNs, users often behave differently in terms of active level. For example, some active users may check in or leave tips very frequently while other users may be inactive and report less digital traces. In our dataset, the average number of observed fine-grained preference per user is 51.40. Therefore, we split each test set into two subsets: low active users (observed preference number < 50) and high active users (observed preference number $>= 50$). Moreover, for a given u-k pair, the number of venues with positive preference in the test set (i.e., the ground truth length $|V^+_{u,k}|$) might be different. In order to prove that MT-RTF algorithm achieves consistently good performance, we report separately the results for different $|V^+_{u,k}|$, and the average performance as well. Figure 4.12 illustrates the results for both low active users and high active users.

First, MT-RTF algorithm gets consistent results (MAS around 0.7) for venue ranking with different length $|V^+_{u,k}|$. The result confirms that MT-RTF preforms well regardless the ground truth length. Moreover, with regard to user active level, a slight improvement could be observed for the high active users (average MAS is 0.7308) comparing with the low active users (average MAS is 0.6840). This observation implies that the more activities users have in LBSNs, the better location search experience they can get.

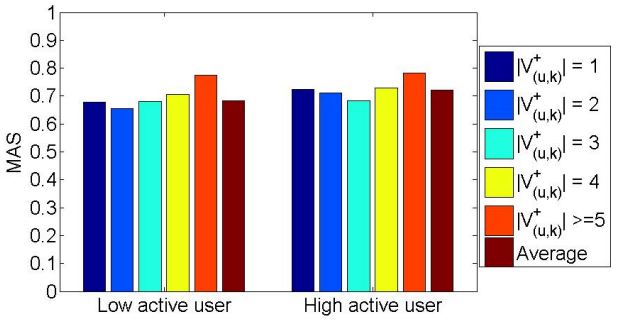


Figure 4.12 – Performance for different types of users.

4.6 Concluding Remarks

In this chapter, by exploring individual preference using heterogeneous user activity data from location centric social media, we propose a sentiment enhanced personalized POI recommendation and search framework. Specifically, we define two types of user preference on POIs, i.e., coarse-grained and fine-grained user preference, and extract them from heterogeneous user activity data, i.e., check-ins and tips. In order to enable personalized location based services, namely, POI recommendation and search, we propose two low-rank approximation based algorithms for user preference prediction tasks. First, by modeling coarse-grained user preference as a matrix, we formulate the user preference prediction problem as a matrix factorization problem, and propose a location based matrix factorization algorithm considering both social network and inter-venue influence. Second, by modeling fine-grained user preference as a tensor, we formulate the personalized ranking problem as a ranking tensor factorization problem, and propose a multi-tuple based ranking tensor factorization algorithm considering both positive and negative user preference in the factorization process. To validate the proposed framework, we experimentally evaluate its effectiveness of rendering personalized services. The results show that our framework cannot only subtly capture user preference with different granularity from user activity data in LBSNs, but also outperform state-of-the-art personalization approaches in user preference prediction tasks.

Although personalization of location based services can improve user experience, it still faces a major challenge, i.e., privacy. Many users would have privacy concerns in using location centric social media [39]. However, privacy and personalization are somehow contradictory. More user data exposed to service providers usually implies better understanding about user preference, and thus leads to better personalization performance. In current literature, researchers have started to investigate the trade-off between privacy pro-

tection and personalization performance [98, 121]. In the future, we also plan to explore more on this issue. In addition, studying user preference in LBSNs can be broadened in other directions. First, as user preference has been well studied in online merchant services, the transfer learning technique may be adopted to augment user preference in LBSNs. Second, with the continuous increase of user activity data, it is necessary to explore new ways of accommodating these accumulated data from LBSNs to enable scalable personalized location based services. Third, since user activities usually show obvious spatial temporal regularity, we will explore more about their spatial temporal patterns and the application in enabling context-aware personalization, which will be presented in the next chapter.

This work was originally published in [149–151].

Chapter 5

Modeling Spatial-Temporal User Activity Patterns

Contents

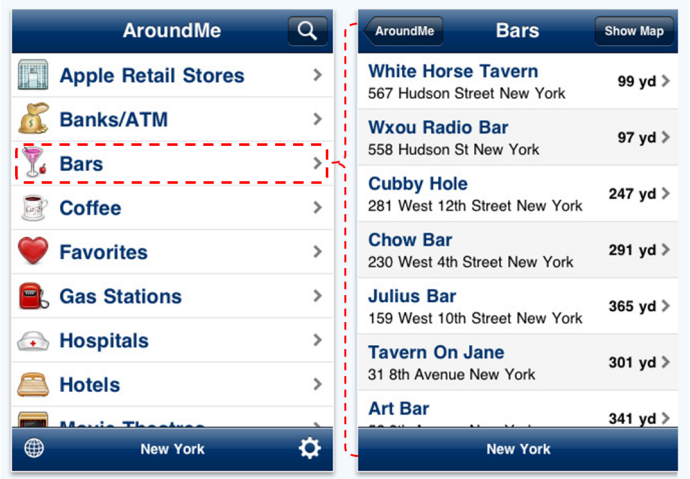


Figure 5.1 – Graphical user interface of AroundMe.

5.1 Introduction

In location centric social media, users leave their spatial-temporal digital footprints, such as check-ins. This data brings an unprecedented opportunity to understand the spatial and temporal regularity of user activity. Typically, a check-in indicates that a user visited a POI at a certain time. Along with POI categories that are often associated with user activities, we can semantically characterize the activities of a user in a place. For example, a user, Jane, is having French food (i.e., being at a French restaurant) at [40.7586, -73.9791] at 21:10 on Friday. By mining these activity records, we are able to understand user spatial temporal activity preference which can then enable various location based applications. The most straightforward application is POI recommendation. For example, knowing Jane is currently interested in going to a Chinese restaurant, a recommendation of Chinese food in a nearby POI would be persuasive. Moreover, knowing a group of users' activity preference, real-time group-oriented advertisement can be better enabled. For example, a clothing store is offering a group discount, if we know five users in the area are interested in the clothing store, an invitation to them would be welcome by both business owners and customers.

In this chapter, we try to answer the following question: "which activity is a mobile user interested in given her current context, including time and location?" Concretely, we aim at modeling user spatial temporal activity preference by leveraging user generated digital footprints in LBSNs. In LBSNs, user generated digital footprints usually contain rich semantic information on their activities. For example, a user's check-in at a French restaurant probably means that the user is having French food there. In current literature [80,99,111,153], POI categories are often considered as the representation of user activities.

In addition, as shown in Figure 5.1, a mobile application called "AroundMe" [1] implements a two-step user interface for users to explore the nearby places, which first lets them select activity category and then shows the specific POIs. Modeling user spatial temporal activity preference is able to improve the user experience of location based services. Taking the "AroundMe" application as an example, given a user's current GPS coordinates and current time, if we infer that the user is interested in going to a bar, the bar category should appear at the very top of the category list in the application. However, modeling user spatial temporal activity preference from user check-ins in LBSNs is not trivial.

— First, since the check-in data is usually sparse and is represented as user-location-time-activity quadruples that contains four data dimensions, it is difficult and complicated to directly discover the regularity from such sparse high-dimensional data.

— Second, to consider spatial dimension, the existing works usually segment a city into disjoint grid cells and discretely infer user preference in individual cells such as in [153]. This method may cause inaccuracy due to the discretization process. For example, when a user is located at the border of two adjacent cells, a movement with a very short distance may incur the change of cells and cause different preference inference results. However, due to the continuity of location dimension, it is not easy to model user spatial activity preference in a continuous manner.

— Third, different from the continuously sampled user activity data, check-ins are user voluntarily reported activities. Most of users do not regularly perform check-ins, due to the reasons such as lack of time and privacy concern, etc. Therefore, check-ins in LBSNs usually suffer from a data sparsity problem, which causes difficulties in modeling user activity preference.

Aiming at resolving the above research issues, we develop a novel user Spatial Temporal Activity Preference (STAP) model. First, in order to reduce the problem complexity, we separately consider the spatial and temporal characteristics of user activity preference in LBSNs. Second, to capture the spatial features, instead of segmenting a city into grid cells, we build Personal Functional Regions for each user using her check-ins, which can then be used to infer ones' spatial activity preference. Third, to resolve the data sparsity problem in capturing temporal features, we exploit other similar users' activities and collaboratively build one's temporal activity preference model. Finally, a context-aware fusion framework is proposed to combine them together.

In the following sections, we first describe two unique spatial and temporal features of user activities that we use to build individual spatial and temporal models, and then present our contribution in modeling user spatial temporal activity preference.

1. http://www.aroundmeapp.com/

Figure 5.2 – Spatial distribution of three users' activities in Manhattan (Check-ins of three different users are plotted in red, green, blue colors, respectively.)

5.1.1 Observations from A Study of User Activities

In order to build a hybrid spatial temporal user preference model, we would like to consider the spatial and temporal features of user activity separately. To this end, we collect and investigate a check-in dataset from a well-known LBSNs Foursquare, and obtain the following observations:

Spatial specificity. Users' activities in LBSNs often show strong preference bias in their frequented regions. In other words, users only conduct a few types of activities (i.e., visit POIs of a few categories) in each of their frequented regions. Figure 5.2 shows check-ins of three New York users (represented by red, green, blue colors) in Manhattan in our dataset. First, we observe clearly that most of a user's check-ins happen in certain geographic areas, as plotted in the Figure. Such observation indicates that check-in behaviors have strong geographic preference and different users usually have their own frequented regions. Second, by investigating users' activities in their frequented regions, we discover that their activities are often limited to a few categories for majority of their frequented regions. We show the dominant activities in one frequented region of each user in Figure 5.2.

Temporal correlation. While users' activities in LBSNs can reflect their temporal activity preference, due to the sparsity of user check-ins, individual's digital footprints cannot well characterize a user's temporal activity preference. For example, if we consider weekly activity preference with hour granularity (i.e., 168 hours in a week), there are 103 hours on average for each user that we did not observe any activity in our dataset. However, we observe that some users' temporal activity preference may be very similar, which naturally fits the underlying assumption of collaborative filtering techniques, i.e.,

Figure 5.3 – Activity category tag clouds of five New York users who share similar activity patterns. (Larger font size implies a higher frequency, and vice versa.)

users who have similar temporal preference on some activities are likely to have similar temporal preference on others. For example, Figure 5.3 illustrates the activity category tag clouds of five similar users in different time slots in New York who share similar activity patterns. The selection of these users as a group is based on the community detection method proposed in [143]. We observe that they usually go to a coffee shop or a burger joint between 13:00 and 14:00 of weekday, stay at a bar between 21:00 and 22:00 on Friday, and go to gym or outdoor places between 16:00 and 17:00 on weekend.

5.1.2 Our Contribution: STAP Model

Based on the previous observations, we introduce STAP model. Concretely, it first separately considers spatial and temporal features of user activity preference and then combines them together using a context-aware fusion framework.

5.1.2.1 Capturing Spatial Feature

Due to the continuity of location and sparsity of check-ins, it is impossible to observe users' activities at all locations. However, spatial specificity suggests that users usually have activity preference bias in their frequented regions. Therefore, we may first try to estimate users' activity preference in their frequented regions, and then infer users' activity preference on their unvisited locations using interpolation methods.

The research challenge here is to discover all those regions for each user and quantitatively measure such preference bias, and then continuously infer user spatial activity preference. Intuitively, for a specific user, a good region should be an area where the user frequently visits and has strong preference bias (i.e., among various categories of activity available there, the user often conducts a few categories of activities). Because equally conducting all categories of activities in that area means the user has no obvious activity preference there. Therefore, we propose Personal Functional Region (PFR). Concretely,

we first define a frequented region of a user as an area with a center and a radius where the user performs more than a certain percentage of all her check-ins. We then measure her activity preference bias in such a region using an entropy-based measure called ratio of preference bias that describes how deterministic the user's activities are in that area. Finally, we define Personal Functional Region as a user frequented region where the user has strong preference bias on certain activities (i.e., the ratio of preference bias is higher than a threshold). Knowing the PFRs of a user, we can then continuously infer her spatial activity preference on her unvisited locations using interpolation methods.

To find out one's PFRs, due to the continuity of location, it is impossible to exhaustively enumerate all regions and identify PFRs. Therefore, we propose a greedy clustering based approach to discover PFRs using user's historical check-ins. The basic idea is to start from one's most frequently visited POI and its neighboring area, because such a region are most likely to be a frequented region, which is a necessary condition for a PFR. Specifically, it first scans from the most checked POI and considers the POI's GPS coordinates as a region center. By evaluating the nearby (i.e., within a certain radius) user activity frequency and ratio of activity preference bias, it then determines whether such a region is a PFR using a threshold based criterion.

5.1.2.2 Capturing Temporal Feature

Temporal correlation shows that some users may have similar temporal activity patterns. It suggests that user temporal activity preference can be collaboratively inferred. Collaborative filtering techniques are widely adopted when tackling sparse data, especially in recommendation systems to predict user preference using limited and sparse user historical data. To collaboratively build user temporal activity preference model using sparse check-ins, we need to first find the latent correlation, i.e., users with the similar temporal activity preference as shown in Figure 5.3, and then infer one's temporal activity preference with the help of the preference of similar users.

The research challenge here is how to discover and leverage the latent correlation. Instead of manually identifying and explicitly describing such correlations, low-rank approximation techniques, such as matrix/tensor factorization, are usually adopted to discover the latent correlation on such multi-facet data. In this work, we use a three-way tensor to model user temporal activities (i.e., user-time-activity), where we consider the weekly user activity pattern with hour granularity because users often exhibit different daily patterns in a week and it is meaningless to measure user activity duration in seconds or minutes in our case. Using non-negative tensor factorization techniques, we are able to discover the latent correlation between user, time and activity factors. By recovering a tensor from these factors, we obtain the approximated non-negative preference measure for each user-

time-activity triplet.

Moreover, compared to the continuously sampled user activity data, check-ins are user voluntarily reported activities. Such a property implies that the sequential patterns of check-ins are not reliable as shown in [47]. Hence, we ignore the sequential patterns of user activities in STAP model and we later consider sequential pattern mining approaches as baselines in evaluation.

5.1.2.3 Fusion of Spatial and Temporal Feature

Due to the complexity of handling the user-location-time-activity data directly, we separately consider spatial and temporal features of user activities and then propose a context-aware fusion framework integrated in STAP to infer user activity preference. The existing works mainly leverage weighted average methods. However, since the ability of spatial and temporal models varies depending on users' contexts (i.e., time and locations), it is difficult to identify the optimal weights for fusion across different contexts. Therefore, we propose to simply use 1/0 weight to combine spatial and temporal models together dynamically according to users' current contexts. Concretely, we first calculate the activity preference inference success rate of both spatial and temporal models on a validation dataset for different contexts. When inferring user activity preference, we then choose the better model for the given context by comparing the success rate of the two models.

We experimentally evaluate STAP using three check-in datasets collected from two LBSN services, i.e., Foursquare and Gowalla [2]. The experiment results show that the STAP model achieves consistently good performance with all three datasets and outperforms various baseline approaches, which verifies the generality and advantages of our solution in modeling spatial-temporal activity preference with sparse check-in data.

5.2 Problem Definition

The objective of this work is to model and infer user spatial temporal activity preference. In LBSNs, users visit diverse categories of POIs. Since POI categories usually imply the activities that users conduct there, we consider POI categories as user activities. Therefore, we are interested in the following problem: given a set of users' historical behaviors, i.e., check-ins, the objective is to infer their interest in activities (visiting certain categories of venues) for a given time, around the current geo-location.

Formally, given a set of users U and a set of venues V related to a set of categories C, each venue belongs to a category c, where $c \in C$. Let C_l^d denote the existing location categories within d km from the Geo-location l (represented by GPS coordinates). Each

2. urlhttp://blog.gowalla.com/

Table 5.1 – Notation.

Symbol	Description
U	set of users
u	a user, $u \in U$
V	set of venues
$\mathcal{A}_u$	check-in activities of user u
$\mathcal{A}_{u,r}$	check-in activities of user u in region r
C	set of POI categories
C_l^d	existing activity categories within center l and radius d
$C_{u,r}$	user u conducted activity categories in region r_u
c	an activity category, $c \in C$
l	GPS coordinates
T	set of time slots
t	a time slot, $t \in T$
r	a region with center l and radius d
r_u	a personal functional region of user u
$\mathcal{R}_u$	personal functional regions of user u
$\psi_{u,r}$	user u's activity distribution in region r
$\Psi_{u,l}$	spatial activity preference of user u at location l
$\Psi_{u,t}$	temporal activity preference of user u at time t
$\Psi_{l,t}$	spatial temporal activity preference of user u at time t and location l

check-in can then be represented by a quadruple (u, l, c, t), representing the user u conduct activity c, at time t when user's position is l. Let $\mathcal{A}_u$ denote the check-in activities of the user u. The problem of modeling user spatial temporal activity preference can then be formulated as: Knowing the historical activities of users U, i.e.,$\{\mathcal{A}_u | u \in U\}$, given a user u whose current position is l at current time t, our aim is to infer u's preference in visiting the nearby venue categories C_l^d. The notations used in this chapter are summarized in Table 5.1.

5.3 Modeling Spatial Patterns of User Activity

In order to model user spatial activity preference in a continuous manner, we propose Personal Functional Region by considering the spatial specificity feature of user activity. The concept of urban functional region [6] has been studied for years. For example, Yuan et al. [159] proposed a framework to discover and semantically annotate urban functional regions using human mobility and POIs in a city. Kurashima et al. [70] proposed a method called Geo Topic Model to discover different activity areas in a city and user's interest for the purpose of location recommendation. Although these works managed to find out the common functional regions in a city, the empirical study shows that different users often have different activity preference in the same area. Based on this observation, we propose PFR which is able to capture individual's spatial activity preference. In this section, we first give the definition of Personal Functional Region and then propose a PFR discovery

algorithm by mining users' historical activities. Finally, we show how to infer user activity preference using PFRs.

5.3.1 Personal Functional Regions

The definition of Personal Functional Region is based on the spatial specificity feature of user activity preference, which shows that users usually perform certain specific activities in their frequented regions. Therefore, in the following, we first define user frequented regions and then define the ratio of preference bias in a frequented region to quantitatively characterize user activity preference bias. Afterwards, we give the definition of Personal Functional Regions.

Definition 1 (Frequented Region)**.** *A region is a geographical area with a center l and a radius d. Region r is a Frequented Region of user u if and only if user u has performed more than s_{freq} of her total check-ins, i.e., the fraction of u's check-ins activities in r is greater than or equal to the threshold s_{freq}.*

$$freq = \frac{\mathcal{A}_{u,r}}{\mathcal{A}_u} \geq s_{freq} \tag{5.1}$$

In this definition, l and d determine the location and the size of the region. The threshold s_{freq} determines the lower bound of the frequency u visits r. Note that we use circular region for frequented region representation due to its simplicity. However, in urban planning community, the most popular methods of describing functional regions in a city are based on its road segmentation and are usually non-overlapped as in [159]. However, based on the geographical distribution of individual's activities, PRFs may have more complex geographical representation, such as polygonal areas. We will investigate different geographical representations of PFRs in our future work.

Functional regions in a city are usually characterized by the distribution of venue categories. For example, a region with lots of stores and restaurants is likely a commercial center; a region with many monuments and historical sites is probably a tourism spot. To describe individual's functional regions, we need to quantitatively measure users' activity diversity in their frequented regions. Since users' activities in their frequented regions usually fall into a few categories rather than all the existing categories C_l^d, we are inspired by the definition of relative redundancy in information theory and propose ratio of preference bias to characterize one's activity preference. Specifically, we measure how deterministic one's activity is in a frequented region r by calculating the difference between the entropy of the users' actual activity distribution and the maximum entropy of the activity distribution in r. The maximum entropy of the activity distribution $H_{max}(|C_l^d|)$ is calculated as

follows:

$$H_{max}(|C_l^d|) = \log_2 |C_l^d| \tag{5.2}$$

It corresponds to the situation that a user u visits all the location categories C_l^d in the region r equally, which means that u does not have obvious preference bias on activities in r. With this reference, we define the ratio of preference bias as follows:

Definition 2 (Ratio of Preference Bias)**.** *For a user u and one of her frequented regions r, the ratio of preference bias measures the fractional difference between the entropy of u's activity distribution $\psi_{u,r}$ in r and the maximum entropy of the activity distribution in r, which is calculated as follows:*

$$ratio_{PB} = 1 - \frac{H(\psi_{u,r})}{H_{max}(|C_l^d|)} \tag{5.3}$$

Higher value of $ratio_{PB}$ implies the stronger activity preference bias of u in r, and vice versa. Since the PFR of a user should be an area where the user has strong activity preference bias, we thus define Personal Functional Region as :

Definition 3 (Personal Functional Region)**.** *A Personal Functional Region (PFR) r_u for user u is a user frequented region where u's activities have higher ratio of preference bias, i.e., $ratio_{PB}$ is greater than or equal to a threshold $s_{ratio_{PB}}$.*

A PFR r_u is then represented by the center l, radius d, u's activity distribution $\psi_{u,r}$ and ratio of preference bias $ratio_{PB}$. As the threshold $s_{ratio_{PB}}$ denotes the lower bound of user activity preference bias in PFRs, we need to identify the boundary of $ratio_{PB}$ in order to properly choose the threshold $s_{ratio_{PB}}$. We show in Proposition **??** that $ratio_{PB}$ is actually bounded to $[0, 1]$.

Proposition 1. *Given any u and any of her frequented region r, the ratio of preference bias $ratio_{PB} \in [0, 1]$.*

The proof of Proposition 1 can be found in Appendix A.1. We now use an example to show how to discover a user's PFR by computing the $ratio_{PB}$ as follows:

Example 1. *A "nightlife region" for a user is shown in Figure 5.4. This user visits mainly "Sports bars" and "Jazz bars". There are totally 20 categories ($|C_l^d| = 20$) in this region. The ratio of preference bias is then calculated as follows:*

$$H(\psi_{u,r}) = -\sum_{c_i \in \psi_{r_u}} p(c_i) \log_2 P(c_i) = 1.78 \tag{5.4}$$

Nightlife Region

Category	Check-in count	Probability
Sports bars	15	0.5
Jazz bars	9	0.3
French Restaurant	3	0.1
Fast food restaurant	2	0.0667
Movie theater	1	0.0333

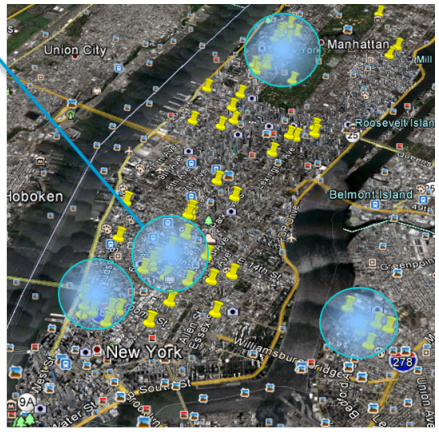


Figure 5.4 – An example of a "nightlife" PFR of a user.

$$H_{max}(|C_l^d|) = \log_2 |C_l^d| = 4.32 \tag{5.5}$$

$$ratio_{PB} = 1 - \frac{H(\psi_{u,r})}{H_{max}(|C_l^d|)} = 0.59 \tag{5.6}$$

5.3.2 PFR Discovery Algorithm

According to above definitions, we need to determine four parameters to specify a PFR based on users' historical activities. Among these four parameters, l refers to the center location of the region; d decides the size of PFR; s_{freq} shows how active a user is in her PFRs; $s_{ratio_{PB}}$ represents the user's activity preference bias degree in her PFRs.

The basic idea of efficiently discovering PFRs is to start from one's most frequently visited POI and its surrounding area, because such a region is most likely to be a frequented region, which is a necessary condition for being a PFR. Therefore, we propose a greedy clustering algorithm to discover PFRs from one's check-in activities as shown in Algorithm 5.2. Concretely, given a user's historical check-ins, we first scan from the most checked venue and consider all the visited venues whose distance is less than d kilometers from the selected venue as a region r (Line 1-6). When calculating the visiting frequency, an activity can only be counted once. We use $\mathcal{A}_{rest}$ to denote the un-counted check-ins. If this region is a user's frequented region (i.e., visiting frequency $freq$ is equal or higher than the threshold s_{freq}), we calculate ratio of preference bias $ratio_{PB}$ (Line 7-10). If $ratio_{PB}$ is equal or higher than the threshold $s_{ratio_{PB}}$, we choose r as a PFR for the user (Line 11-12) and remove the counted check-ins $\mathcal{A}_{u,r}$ from $\mathcal{A}_{rest}$ (Line 13). After examining each check-in venue in all the regions, we get a set of PFRs $\mathcal{R}_u$ for the user u.

Algorithm 5.2 PFR Discovery Algorithm

Input: User u's check-ins $\mathcal{A}_u$ and parameters $d, s_{freq}, s_{ratio_{PB}}$

Sort $\mathcal{A}_u$ in descend order according to visiting frequency
Initialize remainder check-in set, $\mathcal{A}_{rest} = \mathcal{A}_u$
Initialize user u's PRF set, $\mathcal{R}_u = \emptyset$
for $v \in \mathcal{A}_u$ **do**
    if $v \in \mathcal{A}_{rest}$ **then**
        Select r with center $v.l$ and radius d
        Find check-ins $\mathcal{A}_{rest}$ in r denoted as $\mathcal{A}_{u,r}$
        Calculate $freq$ in the region r
        if $freq \geq s_{freq}$ **then**
            Calculate $ratio_{PB}$ in r based on $\psi_{u,r}$
            if $ratio_{PB} \geq s_{ratio_{PB}}$ **then**
                Add r in $\mathcal{R}_u$ with $l, d, \psi_{u,r}, ratio_{PB}$
                Remove $\mathcal{A}_{u,r}$ from $\mathcal{A}_{rest}$
            end if
        end if
    end if
end for
return $\mathcal{R}_u$

5.3.3 Spatial Preference Inference Using PFRs

After discovering users' PFRs, the next issue turns to inferring user activity preference using PFRs. Knowing user u's current location l, we first estimate the preference influence of individual PFRs and then combine activity preference distribution $\psi_{u,r}$ of all PRFs using weighted average methods. Some previous works [31, 33, 103] have studied the probability of location visiting w.r.t. the travel distance, and found that it is inversely proportional. We advocate for their finding and based on that, we propose the following weight function:

$$w_{l,r_u} = \begin{cases} d^{-1}, \ if \ \ d_{l,r_u} \leq d \\ d_{l,r_u}^{-1}, \ if \ \ d_{l,r_u} > d \end{cases} \tag{5.7}$$

Specifically, for the PFRs whose distance d_{l,r_u} from l is less than or equal to the radius d of r_u, the user is currently in these PFRs and we consider their influence equally. For other PFRs whose distance d_{l,r_u} from l is greater than the radius d of r_u, their influence is proportional to d_{l,r_u}^{-1}. Therefore, the spatial activity preference $\Psi_{u,l}$ of user u at location l can be calculated as follows:

$$\Psi_{u,l} = \sum_{r_u \in \mathcal{R}_u} \psi_{u,r_u} \cdot w_{l,r_u} \tag{5.8}$$

5.4 Modeling Temporal Patterns of User Activity

Due to the sparsity of check-in data and the temporal correlation of user activity preference, we exploit other similar user's activities and collaboratively build a user's temporal activity preference. Concretely, we first model user temporal activities using a three-way tensor and then leverage tensor factorization techniques to decompose the tensor into three factors, i.e., user, time and activity factors. By recovering a tensor using these factors, we obtain the preference measure for each user-time-activity triplet. In order to avoid the negative value in the recovered tensor which is meaningless for preference measure, we add non-negative constraint into the factorization process. The non-negative constraint can help to make the results interpretable [76] as probability. In the following, we first present tensor factorization model, and then explain how to infer user temporal activity preference using non-negative tensor factorization techniques.

5.4.1 Tensor Factorization Model

In this work, we build a user-time-activity tensor based on users' historical check-ins. Since we consider venue categories as user activity categories, the tensor is denoted as *u-t-c*, (i.e., user-time-category). Tensor factorization techniques intend to decompose such a tensor into multiple factors. Let $\hat{U}$, $\hat{T}$ and $\hat{C}$ denote the user, time and activity category feature matrices, with size of $|U| * f$, $|T| * f$ and $|C| * f$, respectively. In a sense, these matrices comprise computerized groups of user, time and activity dimension according to users' activities modeled by tensor. For example, a feature dimension for user matrix measures how much a user likes a certain group of temporal activities on the corresponding time and activity feature dimension. Note that f is called latent space dimension (or factorization dimension) which is the most important parameter in tensor factorization. It controls the number of features involved in the factorization process. The decomposition is formulated as follows:

$$\hat{Y} = \hat{O} \times_U \hat{U} \times_T \hat{T} \times_C \hat{C} \tag{5.9}$$

where $\times_n$ is the mode-n tensor product with matrix. The core tensor $\hat{O}$ with dimension $f * f * f$ handles the correlation among different factors. The value of each element in $\hat{Y}$ is calculated as:

$$\hat{y}_{u,t,c} = \sum_{\hat{u}} \sum_{\hat{t}} \sum_{\hat{c}} \hat{o}_{\tilde{u},\tilde{t},\tilde{c}} \cdot \hat{u}_{u,\tilde{u}} \cdot \hat{t}_{t,\tilde{t}} \cdot \hat{c}_{c,\tilde{c}} \tag{5.10}$$

where $\tilde{u}, \tilde{t}, \tilde{c} \in \{1, ..., f\}$ are indices of latent features. This model is called Tucker decomposition model [137]. For simplicity, we assume that the core tensor $\hat{O}$ is a diagonal

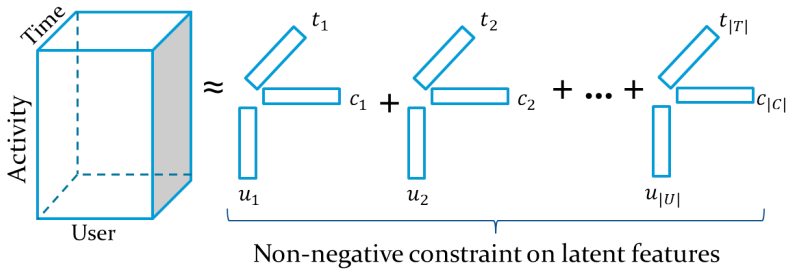


Figure 5.5 – Non-negative tensor factorization using Canonical decomposition model.

tensor:

$$\hat{c}_{\tilde{u},\tilde{t},\tilde{c}} = \begin{cases} 1, \ if \quad \tilde{u} = \tilde{t} = \tilde{c} \\ 0, \ else \end{cases} \tag{5.11}$$

We then obtain Canonical decomposition model with each element calculated as:

$$\hat{y}_{u,t,c} = \sum_{\tilde{f}} \hat{u}_{u,\tilde{f}} \cdot \hat{t}_{t,\tilde{f}} \cdot \hat{c}_{c,\tilde{f}} \tag{5.12}$$

where $\tilde{f} \in \{1, ..., f\}$ is the index of latent space. In this work, we adopt non-negative tensor factorization using Canonical decomposition model which can be efficiently calculated within relatively short running time. Figure 5.5 illustrates the factorization model. We decompose a tensor into three factors (i.e., $\hat{U}$, $\hat{T}$ and $\hat{C}$) and try to optimize the loss function between the recovered tensor $\hat{Y}$ and the original u-t-c tensor.

5.4.2 Temporal Preference Inference

We adopt the non-negative tensor factorization implementation [3] in [66]. It adds non-negative constraint to Alternative Least Square based tensor factorization algorithms and uses Canonical decomposition model.

By recovering $\hat{Y}$ from $\hat{U}$, $\hat{T}$ and $\hat{C}$ using Equation 5.12, we obtain a non-negative tensor describing users' temporal activity preference. In order to infer user u's preference (probability of conducting an activity) at time t, we normalize $\hat{Y}$ as follows:

$$\sum_{c=1}^{|C|} \hat{y}_{u,t,c} = 1, \ \forall u \in U \ and \ \ \forall t \in T \tag{5.13}$$

For the given u and t, the sum of all activities' preference measure (i.e., probability) is normalized to one. This is for the later fusion with spatial activity preference which

3. https://sites.google.com/site/jingukim/home#ntfcode

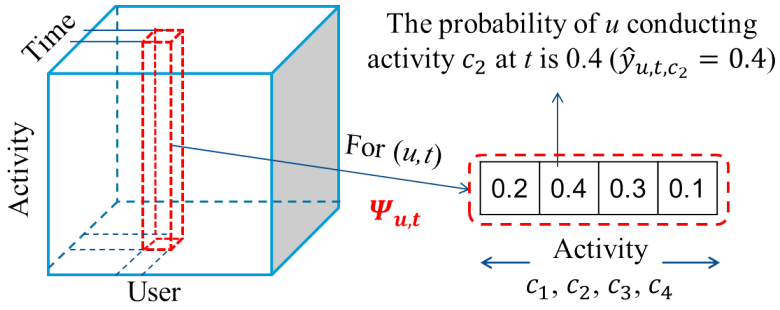


Figure 5.6 – Temporal activity preference inference.

is represented by probability in the value range of $[0, 1]$. Figure 5.6 shows an example with four activity categories, where the normalized preference measure can be regarded as the probability that u conducts activity c at time t. Thus, we obtain temporal activity preference $\Psi_{u,t}$ as follows:

$$\Psi_{u,t} = \{\hat{y}_{u,t,c} | c \in C\} \tag{5.14}$$

5.5 Context-aware Fusion Framework

Given one's spatial and temporal activity preference, i.e., $\Psi_{u,l}$ and $\Psi_{u,t}$, the fusion framework tries to combine them together to obtain the user spatial temporal activity preference. The most straightforward approach is to merge them using two static weights. Since the performance of spatial and temporal models varies over time and locations, the simple weighted average cannot always get the better one of the two models (later proved in evaluation). However, it is difficult to dynamically assign the two weights according to the user context. Therefore, by conducting the study on a validation dataset, we simply select the model with higher accuracy for activity preference inference according to a user's current context (i.e., location l and time t). Therefore, we propose a context-aware fusion framework to take advantage of both spatial and social models. Specifically, we first define the success rate of a preference model as the frequency of correct inference for the Top 1 activity. Then, for each user, we use two matrices to calculate the success rate of both spatial and temporal models on different contexts using a validation dataset. When inferring user activity preference, the model with higher success rate is used.

5.5.1 Success Rate Calculation of Preference Model

The objective of calculating success rate is to get the inference accuracy of both preference models under different contexts, i.e., time and involved PFRs. Let M_{tem} and M_{spa}

Algorithm 5.3 Context-aware success rate calculation

Input: User u's spatial and temporal preference distribution $\Psi_{u,l}$ and $\Psi_{u,t}$, PRFs $\mathcal{R}_u$, activities in validation dataset $\mathcal{A}_{u,valid}$

1: Initialize M_{tem} and M_{spa} with 0
2: **for** $v \in \mathcal{A}_{u,valid}$ **do**
3: Get the Top 1 activity c_l based on $\Psi_{u,l}$
4: Get the Top 1 activity c_t based on $\Psi_{u,t}$
5: Find local PFRs $R_{u,local} = \{r_u \in R_u | d_{l,r_u} \leq d\}$
6: **if** $R_{u,local}$ is not empty **then**
7: **if** c_l equals user actual activity $v.c$ **then**
8: Augment $M_{spa}(v.t, R_{u,local})$ by 1
9: **end if**
10: **if** c_t equals user actual activity $v.c$ **then**
11: Augment $M_{tem}(v.t, R_{u,local})$ by 1
12: **end if**
13: **end if**
14: **end for**
15: **return** M_{tem} and M_{spa}

denote the matrices of spatial and temporal success rate, respectively. Each row of the matrices corresponds to a time slot t and each column represents one functional region r_u of user u. Algorithm 5.3 shows the process of building M_{tem} and M_{spa}. We first initialize M_{tem} and M_{spa} by assigning each element to 0 (Line 1). For each check-in activity in the validation dataset, we infer u's spatial and temporal activity preference $\Psi_{u,l}$ and $\Psi_{u,t}$ and then get the most probable activity c_l and c_t (Line 2-4). We also get u's local PFRs $R_{u,local} = \{r_u \in R_u | d_{l,r_u} \leq d\}$ where the user is currently in (Line 5). If the user's $R_{u,local}$ is not empty, we get the user's current context, i.e., time $v.t$ and local PFRs $R_{u,nearby}$ (Line 6). Afterwards, if the spatial model infers the correct activity, we augment the success rate in M_{spa} for current context, i.e., time slot and local PFRs, by 1 (Line 7-9). Specifically, $M_{spa}(v.t, R_{u,nearby})$ represents the numbers in $v.t$ row and r_u ($r_u \in R_{u,nearby}$) column(s) of M_{spa}. If the temporal model infers the correct activity, we do the same for M_{tem} (Line 10-12).

5.5.2 Fusion Criterion

Knowing the success rate of each model under different contexts, we choose the model with the higher success rate. Algorithm 5.4 shows this process. Specifically, for a given user u and her context, i.e., time t and location l, we obtain matrices M_{tem} and M_{spa} of u generated by Algorithm 5.3. Then, we find u's local PFRs $R_{u,local}$ based on her current location l (Line 1). If $R_{u,local}$ is not empty, we calculate the overall success rate for both

Algorithm 5.4 Context-aware preference fusion

Input: User u's spatial and temporal preference distribution $\Psi_{u,l}$ and $\Psi_{u,t}$, context t and l, PRFs $\mathcal{R}_u$, success rate matrices M_{tem} and M_{spa}

1: Find local PFRs $R_{u,local} = \{r_u \in R_u | d_{l,r_u} \leq d\}$
2: **if** $R_{u,local}$ is not empty **then**
3: Calculate $rate_{spa}$ according to Equation 5.15
4: Calculate $rate_{tem}$ according to Equation 5.16
5: **if** $rate_{spa} > rate_{tem}$ **then**
6: $\Psi_{u,l,t} = \Psi_{u,l}$
7: **end if**
8: **if** $rate_{spa} < rate_{tem}$ **then**
9: $\Psi_{u,l,t} = \Psi_{u,t}$
10: **end if**
11: **if** $rate_{spa} = rate_{tem}$ **then**
12: $\Psi_{u,l,t}$ = randomly choose one from $\{\Psi_{u,t}, \Psi_{u,l}\}$
13: **end if**
14: **else**
15: $\Psi_{u,l,t} = \Psi_{u,t}$
16: **end if**
17: **return** $\Psi_{u,l,t}$

spatial and temporal models as follows (Line 2-4):

$$rate_{spa} = \sum_{r_u \in R_{u,local}} M_{spa}(t, r_u) \tag{5.15}$$

$$rate_{tem} = \sum_{r_u \in R_{u,local}} M_{tem}(t, r_u) \tag{5.16}$$

We then use the one with higher success rate as final preference distribution (Line 5-10). In case of equality, we randomly choose one from $\Psi_{u,t}$ and $\Psi_{u,l}$ (Line 11-13). If u's local PFR set $R_{u,local}$ is empty, we consider the preference distribution of temporal model as the final result (Line 15), because the spatial model is considered to be unconfident in this case.

5.6 Experimental Evaluation

We evaluate STAP by conducting activity preference inference experiments using three datasets collected from two LBSN services, i.e., Foursquare and Gowalla. In the following, we first present the experiment setting including data collection, evaluation plan and metrics. We then show the impact of parameters on STAP model in order to identify their optimal values. Finally, we present the comparison with baseline approaches in terms of preference inference performance.

Table 5.2 – Dataset statistic.

Dataset	New York (Foursquare)	Tokyo (Foursquare)	New York (Gowalla)
Users	824	1,939	244
Venues	38,336	61,858	9,352
Check-ins	227,428	573,703	85,010
Average number of activity categories per user	38.37	31.39	55.58

5.6.1 Experimental Setting

5.6.1.1 Data Collection

In this work, we use three datasets collected from two LBSN services, i.e., Foursquare and Gowalla, to evaluate our model.

Foursquare Dataset. We use a collection of Foursquare check-ins lasting for about 10 months (from 12 April 2012 to 16 February 2013). We filter out noise and invalid check-ins, and then select active users in two big cities i.e., New York and Tokyo, as experiment dataset. Venues in Foursquare are classified into 9 root categories and 291 sub-categories at the time of data collection. Based on these sub-categories, we manually merge the similar and infrequent venue categories together, resulting in a total of 251 venue sub-categories.

Gowalla Dataset. In order to validate that our approach does not depend on the LBSN services, we also conduct experiments using a dataset from another LBSN service Gowalla (from January 2010 to October 2010), which is extracted from the dataset used in [35]. The data filtering and processing step is similar to that of the Foursquare dataset. We select the check-ins in New York as experiment dataset.

The statistics of the selected datasets are shown in Table 5.2. The tag clouds of user activities on the datasets are illustrated in Figure 5.7. Note that there is no obvious difference between the tag clouds of New York (Foursquare) and New York (Gowalla). We thus only show the results from Foursquare. We observe clearly the cultural differences between the two cities: New York users usually share their activities in bars, gyms, restaurants; while Tokyo users often share their presence at train stations, convenience stores, and Japanese restaurants.

5.6.1.2 Evaluation Plan

In the following experiments, we use the first eight-month check-ins as training dataset to build individual spatial and temporal models. We then use the 9th month check-ins as validation dataset to calculate the success rate of individual models for the context-aware fusion framework. Finally, we use the 10th month check-ins as test dataset for experiments.

(a) New York (Foursquare) (b) Tokyo (Foursquare)

Figure 5.7 – Tag cloud of activity category (Larger font size implies higher frequency, and vice versa.) We only show the tag cloud of New York (Foursquare) dataset because there is no visual difference between it and the tag cloud of New York (Gowalla) dataset.

5.6.1.3 Evaluation Metric

Since our application scenario focuses on recommending activities to users, our primary evaluation objective is to see whether a user's interested activity appears at the top of the returned list. Specifically, for each new check-in in the test dataset, we infer the user's activity preference under the given context and compare it with the user's actual activity. Therefore, we use the Top K accuracy ($Accuracy@K$) as the first evaluation metric, which calculates the percentage of the actual activities appearing at the Top K inferred activities in the test dataset. For the test dataset S, the Top K accuracy is calculated as follows:

$$Accuracy@K = \frac{|\{(u,l,c,t)|c \in P_{u,l,t}(K), (u,l,c,t) \in S\}|}{|S|} \tag{5.17}$$

where $P_{u,l,t}(K)$ is the Top K activities inferred for user u at time t and location l (i.e., the Top K activities in $\Psi_{l,t}$ for user u). Moreover, in activity recommendation scenarios such as AroundMe application (Figure 5.1), a user may scroll the screen to find her interested activity. Therefore, the actual rank of the desired activity also has impact on user experience. In order to evaluate the overall ranking of the inferred activity preference, we use Average Percentile Rank [101] of the actual activity in the inferred activity list as another metric, which is calculated as follows:

$$AveragePercentileRank = \sum_{(u,l,c,t)\in S} \frac{|\Psi_{l,t}| - rank(c) + 1}{|\Psi_{l,t}|} \tag{5.18}$$

where $rank(c)$ is the rank of the actual activity c in the inferred activity preference list. The score of Average Percentile Rank is bounded in $(0, 1]$. The high score of Average Percentile Rank implies that the actual user activities appear on the top of the inferred activity preference list, and vice versa.

In addition, from the activity category perspective, some activities may show stronger spatial temporal regularities than others. For example, going to school or office may show stronger spatial temporal regularities than shopping activities. Therefore, in order to study the performance over different activity categories, we consider our problem as a classification problem with 251 activity categories and calculate precision/recall and the related F1 score for each activity category. For a specific activity category c_i, these metrics are calculated as follows:

$$Precision(c_i) = \frac{|\{(u,l,c_i,t)|P_{u,l,t}(1) = c_i, (u,l,c_i,t) \in S\}|}{|\{(u,l,c,t)|P_{u,l,t}(1) = c_i, (u,l,c,t) \in S\}|} \tag{5.19}$$

$$Recall(c_i) = \frac{|\{(u,l,c_i,t)|P_{u,l,t}(1) = c_i, (u,l,c_i,t) \in S\}|}{|\{(u,l,c_i,t)|(u,l,c_i,t) \in S\}|} \tag{5.20}$$

$$F1Score(c_i) = \frac{2 \cdot Precision(c_i) \cdot Recall(c_i)}{Precision(c_i) + Recall(c_i)} \tag{5.21}$$

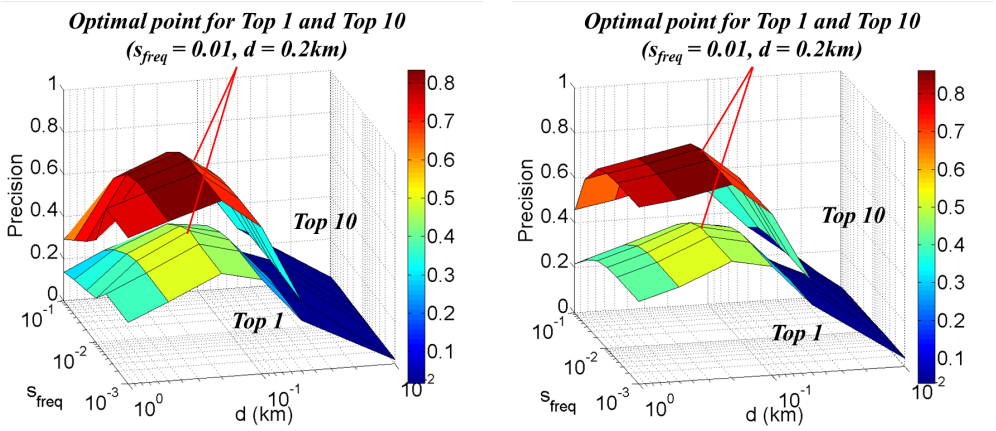
5.6.2 Impact of Parameters on STAP model

The STAP model separately considers spatial and temporal features of user activity preference. From spatial perspective, a user's Personal Functional Regions are determined by four parameters, i.e., l, d, s_{freq} and $s_{ratio_{PB}}$. From temporal perspective, the latent space dimension (or factorization dimension) determines the number of the features involved in the factorization process. Low dimension may result in unsatisfied performance while the high dimension usually implies high runtime complexity. Note that the impact of parameters on STAP model is similar with the New York (Foursquare) and New York (Gowalla) datasets due to the same geographical constraint. We only report the results with New York (Foursquare) dataset and Tokyo (Foursquare) dataset in this section.

5.6.2.1 Spatial Parameter Setting

In order to identify the optimal parameters of Personal Functional Regions for activity preference inferring, we conduct two experiments. First, we show the preference inference accuracy with different parameter combinations of PFRs, i.e., region size d and visiting frequency s_{freq}. Second, by fixing optimal values of the above two parameters, we tune the threshold of preference bias ratio, i.e., $s_{ratio_{PB}}$. Note that we do not study the impact of the PFR centers (i.e., l) in the above experiments because they are automatically determined by the PFR discovery algorithm.

In the first experiment, we set the threshold of preference bias ratio to its lower bound, i.e., $s_{ratio_{PB}} = 0$ according to Proposition 1, which implies that we consider all user frequented regions as PFRs regardless user activity preference bias there. We then plot Top 1



(a) New York (Foursquare) (b) Tokyo (Foursquare)

Figure 5.8 – Parameter tuning of region size d and visiting frequency s_{freq}.

and Top 10 accuracy of activity preference inference by varying s_{freq} within [0.001, 0.005, 0.01, 0.02, 0.05, 0.1], and d within [0.01, 0.05, 0.1, 0.2, 0.5, 1] km.

Figure 5.8 plots the results using the New York (Foursquare) and Tokyo (Foursquare) datasets. For each dataset, we observe a convex surface for preference inference accuracy. We analyze such results as follows.

— *Region Size d.* The small region size (small d) results in bad performance. In this case, users are hardly influenced by their PFRs because their current location rarely belongs to those small PFRs. In contrast, the large region size also generates unsatisfied results because some noisy activities might be included in large PFRs.

— *Threshold of visiting frequency* s_{freq}. The small s_{freq} implies that some detected PFRs might be areas that user occasionally visited. These PFRs are not necessarily reflecting their habitual behaviors and thus cause noise in preference inference. In contrast, the large s_{freq} implies that only highly frequented regions are considered, which cannot fully capture users' habitual behaviors, either.

The optimal value for these parameters ($d = 0.2$km and $s_{freq} = 0.01$) can be identified in Figure 5.8, where the Top 1 and Top 10 activity preference inference accuracy achieves the optimal value. An interesting observation is that the 0.2km optimal radius of PFRs is also in agreement with the optimal urban neighborhood radius identified by urban planning community in [93].

In the second experiment, we fix the optimal d and s_{freq} and decrease the threshold of ratios of preference bias $s_{ratio_{PB}}$ from 1 to 0 with the step of 0.1. Figure 5.9 shows the Top 1, Top 5 and Top 10 inference accuracy. A higher value of $s_{ratio_{PB}}$ implies that we only select

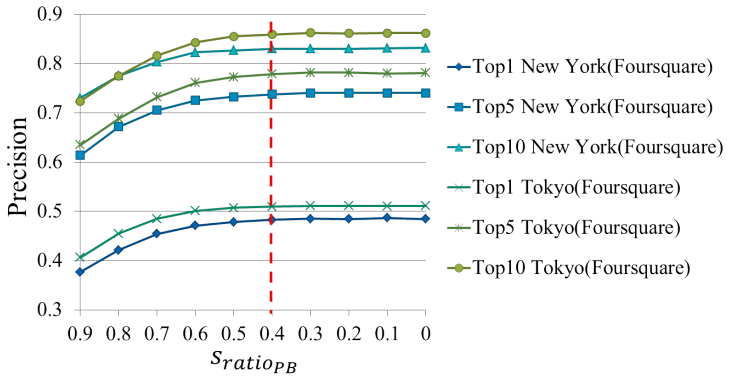


Figure 5.9 – Parameter tuning of threshold of preference bias ratio $s_{ratio_{PB}}$.

PFRs with stronger preference bias. Therefore, some useful PFRs with $ratio_{PB}$ lower than $s_{ratio_{PB}}$ are eliminated so that user spatial activity preference cannot be fully described. With the decreasing threshold $s_{ratio_{PB}}$, more PFRs with relatively lower $ratio_{PB}$ are taken into account. Those PFRs with lower $ratio_{PB}$ has less ability to characterize user spatial activity preference. In the extreme case of $ratio_{PB} = 0$, where users conduct all activities equally, such PFRs cannot help to characterize user spatial activity preference at all. In Figure 5.9, we observe that there is no further improvement for $s_{ratio_{PB}} <= 0.4$ which indicates that the activity preference inference accuracy converges in terms of $s_{ratio_{PB}}$.

In the following, we set the three main parameters of PFRs as $d = 0.2km$, $s_{freq} = 0.01$ and $s_{ratio_{PB}} = 0.4$ for all three datasets.

5.6.2.2 Temporal Parameter Setting

We use the non-negative tensor factorization method to infer user temporal activity preference. The *latent space dimension* controls the number of the features involved in the factorization process. In this experiment, we vary the latent space dimension in the order of 8, 16, 32, 64 and 128. Figure 5.10 reports the comparison results. With the increase of the latent space dimension, the inference accuracy also increases. We observe no significant improvement in inference accuracy for dimension higher than 64, which indicates the convergence in terms of latent space dimension. Hence, in the following experiments, the latent space dimension is set to 64.

5.6.3 Comparison with Baseline Approaches

To evaluate the activity preference inference accuracy of the STAP model, we compare it with the following baseline approaches:

Sequential pattern mining approaches:

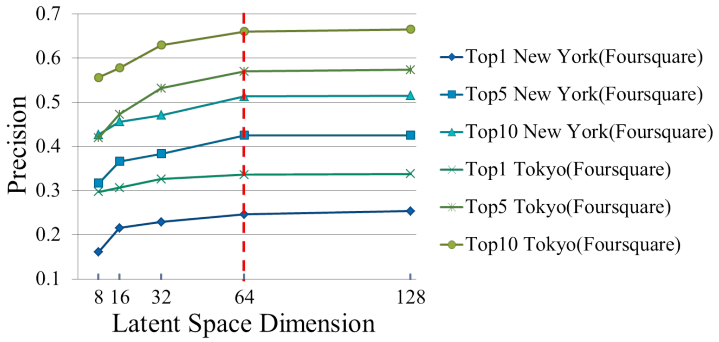


Figure 5.10 – Parameter tuning of latent space dimension.

— *Order-K Markov Model (Markov-K).* People's activities usually follow certain sequential patterns. For example, a user often has coffee after lunch; therefore it would be logical to infer her activity after lunch as having coffee in a coffee shop. Order-K Markov model considers the latest K activities of a user, and searches for the most frequent patterns to predict the next activity. We set K as 1 and 2 in the experiments.

Temporal based approaches:

— *Most Frequent Activity by Time (MFT).* In general, people seem to conduct the same activity in the same time slot, which is usually regarded as a routine activity. For example, a user may have lunch around 12:00 during the weekdays. In this model, one's temporal activity preference is modeled by the distribution of her historical activity categories in each time slot of a week.

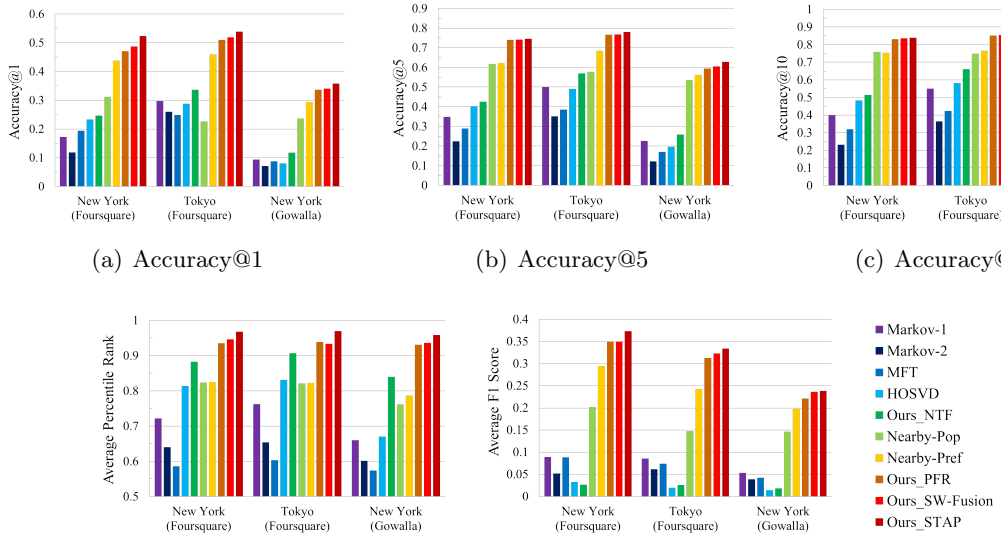
— *High Order Singular Vector Decomposition (HOSVD).* HOSVD [42] is considered as a baseline for tensor factorization approach. It corresponds to the Tucker decomposition optimized for square-loss.

— *Temporal Model of STAP (Ours_NTF).* The temporal activity preference model of STAP uses non-negative tensor factorization (NTF) approaches. It corresponds to the Canonical decomposition optimized for square-loss.

Spatial based approaches:

— *Most Popular Activity Around (Nearby-Pop).* Using one's current location as center, it infers the user's activity preference according to the region's (radius = d) activity popularity. By activity popularity we mean the total number of check-ins for a specific activity category that we observe in the training dataset. It can be regarded as a simple non-personalized functional region based model.

— *Most Preferred Activity Around (Nearby-Pref).* Using one's current location as center, it infers the user's activity preference according to the user's own activity popularity



(a) Accuracy@1 (b) Accuracy@5 (c) Accuracy@10

(d) Average Percentile Rank (e) Average F1 Score

Figure 5.11 – Comparison with baselines under different evaluation metrics.

in the region (radius = d). By one's activity popularity we mean the number of check-ins of the user for a specific activity category that we observe in the training dataset. It can be regarded as a basic personalized model.

— *Spatial Model of STAP (Ours_PFR).* The spatial activity preference model of STAP uses Personal Functional Regions to capture users' spatial activity preference.

Spatial temporal based approaches:

— *Static Weighted Fusion (Ours_SW-Fusion).* The spatial and temporal preference distributions are combined using optimized static weight, i.e., $\Psi_{u,l,t} = \alpha\Psi_{u,t}+(1-\alpha)\Psi_{u,l}$. The optimized α is obtained when inference accuracy is maximized by increasing α from 0 to 1 with the step of 0.1, using the validation dataset. We then find $\alpha = 0.3$ with the New York (Foursquare) dataset and New York (Gowalla), and $\alpha = 0.4$ with the Tokyo dataset (Foursquare).

— *Ours_STAP.* The proposed STAP model uses the context-aware fusion framework.

Figure 5.11 shows the activity preference inference comparison with baselines under different evaluation metrics with the three datasets. We observe that our solution is consistently better than the other baseline approaches. Taking the Top 1 accuracy with the New York (Foursquare) dataset as an example, our solution is 203.54% better than the best sequential pattern mining approach, 124.83% better than the best temporal based approach and 68.09% better than the best spatial based approach. We also conduct the one-tailed

and two-tailed paired t-test over the results. We find that all the p-values are much less than 0.01, which proves that our STAP model is significantly better than the baselines in spatial temporal user activity inference task. In the following, we further analyze the performance of each baseline approach.

First, for sequential pattern mining approaches, both order 1 and order 2 Markov Models obtain unsatisfied results. In LBSNs, users have a choice to share their location information. Therefore, although user activities follow certain sequential patterns in their daily life, user check-ins do not fully contain their daily activities due to privacy concern or lack of time. Moreover, since we consider activities with fine granularity including 251 categories rather than 9 top categories, the large number of categories further aggravates the sparsity issue in the Markov Model when searching for frequent patterns.

Second, for temporal based approaches, tensor factorization methods, i.e., NTF and HOSVD, can better capture user activity preference than the frequency based approach, i.e., MFT. This observation shows that collaborative filtering can efficiently handle the sparse check-in data for user temporal activity preference inference. Furthermore, the improvement of NTF over HOSVD shows the advantage of considering non-negative constraint. The proposed temporal model using NTF can effectively capture the temporal characteristics of user activity preference, particularly for the users whose activities show strong temporal regularities.

Third, spatial based approaches lead to better performance than temporal based methods. This observation shows that the spatial regularity of user activity in LBSNs is more significant than the temporal regularity. Specifically, Nearby-Pref performs better than Nearby-pop baseline due to the consideration of personal preference. The improvement of PFR over Nearby-Pref shows the advantages of eliminating noisy data in capturing spatial features of user activity preference. In other words, the infrequent activities of a user may not actually reflect her preference. The proposed Personal Functional Region can delicately capture the spatial characteristics of user activity preference, particularly for the users whose activities exhibit obvious spatial specificity.

Finally, compared to the static weighted fusion method SW-Fusion, the context-aware fusion framework achieves the best performance. It takes advantage of both spatial and temporal features under varying contexts. An interesting observation is that the improvement of considering the temporal model from merely considering the spatial model is relatively small, which further shows that the importance of spatial features in modeling user activity preferences in LBSNs. Moreover, by comparing the *Accuracy*@1 of the spatial and temporal models, we find that a large number of activities can be correctly inferred by both spatial and temporal models. For example, there are 20.6% of the activities in the test dataset can be correctly inferred by both models for the Top 1 accuracy with the New

York (Foursquare) dataset.

5.6.4 Comparison between Different Datasets

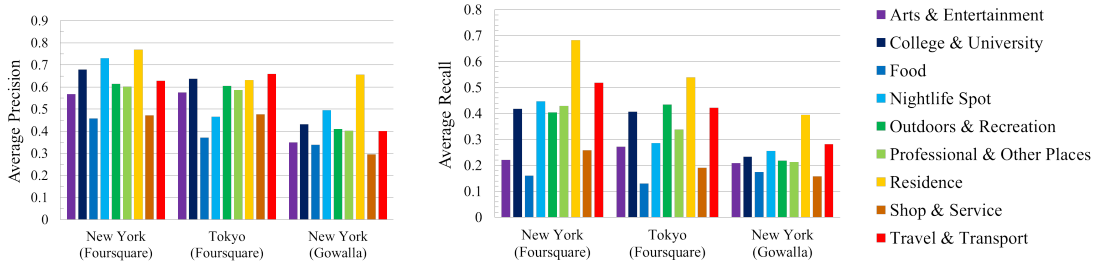
By comparing the results obtained from different datasets, we observe some interesting findings.

First, comparing the results on the New York (Foursquare) and Tokyo (Foursquare) datasets, we find that: a) the accuracy difference between temporal based approaches and spatial based approaches is relatively small with the Tokyo dataset (e.g., 0.34 for Top 1 NTF and 0.51 for Top 1 PFR) than that with the New York dataset (e.g., 0.25 for Top 1 NTF and 0.47 for Top 1 PFR). The most possible explanation is that Tokyo users' activities have stronger temporal regularities than those of New York users; b) the improvement of PFR-based approaches over the non-personalized functional region based approach (Nearby-Pop) is larger with the Tokyo dataset than with the New York dataset, particularly for Top 1 activity inference accuracy. Such an improvement may probably be explained by two reasons: 1) higher density of venues implies higher diversity of nearby activities in Tokyo, which causes less difference for top popular activities; 2) Tokyo users have stronger preference bias in their PFRs, resulting in higher accuracy for PFR-based approaches.

Second, comparing the results on the New York (Foursquare) and New York (Gowalla) datasets, we find that our solution consistently achieves better performance than the baselines. This is due to the fact that users in different LBSNs often exhibit similar spatial-temporal preference activity patterns, which enables us to model the user activity preference over different LBSNs. Furthermore, we find that the performance is slightly better with the New York (Foursquare) dataset than that with the New York (Gowalla) dataset. It can probably be explained by the fact that users in New York (Gowalla) dataset show broader activity preference than that of users in New York (Foursquare) dataset, which makes it more difficult in preference inference task. Specifically, Gowalla users indicate activities in more categories (55.58 on average per user) than Foursquare users (38.37 on average per user).

5.6.5 Comparison between Different Activity Categories

Due to the fact that users' behaviors in some activity categories often show stronger spatial temporal regularities than that in other categories, we investigate such difference by calculating the precision and recall for individual categories using STAP model. Rather than exhaustively listing all the results for the 251 categories, we present the average precision and recall for each of the 9 root-categories in Foursquare, i.e., Arts & Entertainment, College & University, Food, Great Outdoors, Nightlife Spot, Professional & Other Places,



(a) Precision (b) Recall

Figure 5.12 – Comparison between different activity categories using STAP model.

Residence, Shop & Service, Travel & Transport. Figure 5.12 presents the results with the three datasets.

First, in all the three datasets, we observe that categories in Residence and College & University yield good precision and recall, which implies that users in LBSNs exhibit strong spatial temporal regularities in activities like going home and going to school. Intuitively, these activities are usually conducted at regular times and places. We also observe that categories in Shop & Service and Food show relatively lower precision and recall, which implies that users in LBSNs have relatively flexible temporal and spatial preference for shopping or going to a restaurant.

Second, there are also some activity categories yielding different results with different datasets. Specifically, with both New York (Foursquare) and New York (Gowalla) datasets, nightlife activities exhibit high precision and recall, which implies that New York users tend to enjoy their nightlife at regular time and places. In addition, transportation related activities show higher spatial temporal regularities with Tokyo (Foursquare) dataset, which is probably due to the fact that Tokyo users often check in when they are on their daily commute.

5.7 Concluding Remarks

Understanding the spatial temporal patterns of user activity can benefit users by providing them with customized location based services. However, it is difficult to directly tackle such four dimensional data, i.e., user-location-time-activity quadruples, which usually suffers from data sparsity problem. We present STAP, a spatial temporal activity preference model. To reduce the problem complexity, STAP separately considers the spatial and temporal features of user activities by introducing the notion of spatial specificity and temporal correlation. First, spatial specificity suggests that users usually conduct cer-

tain specific activities in their frequented areas. We define Personal Functional Regions to quantitatively measure one's preference bias in her frequented regions and use them to infer spatial activity preference. Second, temporal correlation suggests that users with the similar lifestyle tend to have similar activity preference at the similar time. We resort to tensor factorization techniques to collaboratively build temporal activity preference from the sparse check-in data. Finally, we propose a context-aware fusion framework to make best use of the advantage of both features in activity preference inference. We experimentally evaluate STAP using three datasets collected from two LBSNs, i.e., Foursquare and Gowalla. The experiment results show that the STAP model achieves consistently good performance with all three datasets and outperforms various baseline approaches, which verifies the generality and advantages of our solution in modeling spatial-temporal activity preference with sparse check-in data.

In the future, we plan to broaden this work in several directions. First, since functional regions in urban planning community usually have more complex geographical representations, such as polygonal areas based on the road segmentation in a city, we plan to study different geographical representations of PFRs in order to better characterize user spatial activity preference. In addition, while many research works suggests that the social relationship usually influence user mobility, we will explore more the impact of social network on the spatial temporal user activity patterns. Finally, we may also consider to exploring rich user profile data, such as information on one's homepage or business card [51, 52], and rich context information, such as current weather [74], in order to enable more efficient context-aware location based services.

This work was originally published in [152].

Chapter 6

Exploring Global-scale Nation-wide Collective Behavior

Contents

6.1 Introduction

In the long history of human development, human behavior has been widely studied across various disciplines, such as psychology, biology, sociology and economics, etc [130]. When studying human behavior, we can understand not only individual's behavior, such as one's gestures and facial expressions, but also collective behavior, such as crowd mobility and social movement. In this chapter, we focus on collective human behavior, which can

be defined as the behavior of aggregates whose interaction is affected by some sense that they constitute a group but who do not have procedures for selecting or identifying leaders or members [138]. For example, people in New York city usually go to central business districts for work from residential areas in the morning; French people often go to French restaurants in the evening for dinner while Japanese usually go to bars after work.

However, it is practically difficult to collect large-scale collective behavior. In current literature, traditional collective behavior studies are usually conducted based on some dedicatedly designed experiments [107]. Due to such setting, it is hard to carry out collective behavior experiments on a large population and collect large-scale data.

Fortunately, the increasing popularity of Location Based Social Networks (LBSNs) makes large-scale user behavior data become attainable. In LBSNs, users can share their real time presence with their friends by checking in at POIs. Along with the POI category, we are able to understand the semantic meaning of the check-in activity [152]. For example, a user's check-in in office probably means the user's current activity is working. By interacting with LBSNs, users left a significant volume of check-in data. This data massively implies the physical behavior of users and provides us with an unprecedented opportunity to explore large-scale collective behavior. For example, by analyzing the check-in data across different populations (e.g., people in different countries), we may discover certain behavioral differences between them.

In order to select an appropriate granularity of populations for our study, in this chapter, we focus on collective behavior in individual countries, because countries are usually the subject of inquiry of both politics and economy. For example, the mobility of citizens are usually bounded by the territories of their countries; the "rules of games" (e.g., legal rules and code of ethics) also vary across different countries; various macroeconomic statistics, such as gross domestic product (GDP) and inflation rate, are usually reported with country granularity. There exists also a Web service named "Nation Master" [1] that collects social and economic data by country from various sources and provides different visualization of the data. Figure 6.1 illustrates its screenshot for comparison between two countries (i.e., the United States and Japan).

When studying collective behavior with country granularity, one of the primary tasks is to understand the behavioral differences between countries. For example, when an American would like to travel to Japan for the first time and intends to enjoy a concert there, she may be wondering whether "Japanese people usually go to concert earlier than Americans do?", in order to better plan her trips. To answer such a question, we need to study the traffic patterns (i.e., visiting frequency at different time) of concert halls in the United States and Japan. The collective check-ins in LBSNs massively imply the traffic patterns of

1. http://www.nationmaster.com/

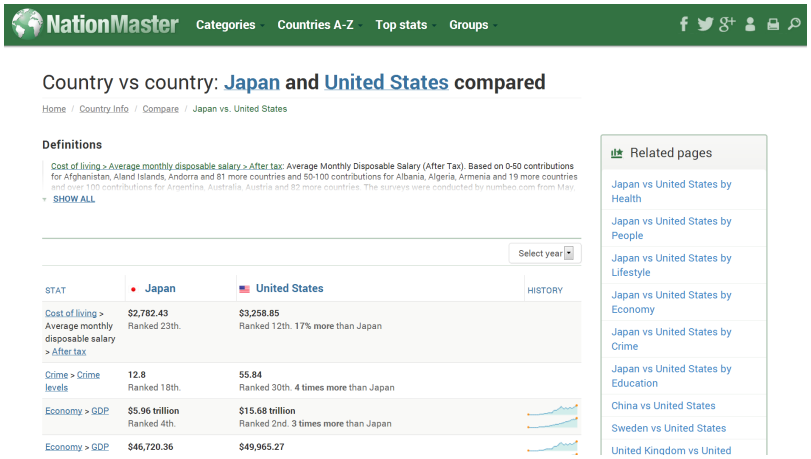


Figure 6.1 – Screenshot of NationMaster (comparison between the United States and Japan).

each POI category. By extracting and comparing such traffic patterns in different countries, we are able to discover their behavioral differences.

In this chapter, we present NationTelescope, a platform that monitors and visualizes large-scale nation-wide collective behavior in LBSNs, and supports the collective behavior comparison between countries. Specifically, it incorporates three unique features.

— First, as users continuously report their activities (i.e., check-ins) in LBSNs, it collects user behavior data on a global scale via check-in data streams from LBSNs.

— Second, in order to efficiently visualize such large-scale data, it automatically generates data summary (i.e., various statistics of collective behavior) and integrates a map interface to visualize the summarized data using interactive map techniques.

— Third, in order to efficiently identify and visualize behavioral differences between countries, it incorporates a discriminative traffic pattern search method to detect discriminative activities (represented by POI categories) between countries.

By developing a prototype of NationTelescope platform, we evaluate its effectiveness and usability via two case studies and a System Usability Scale (SUS) [22] survey. The results show that the platform can efficiently capture and visualize the collective behavior in countries, and effectively compare collective behavior in different countries. The SUS survey with 18 participants proves the good usability of the platform.

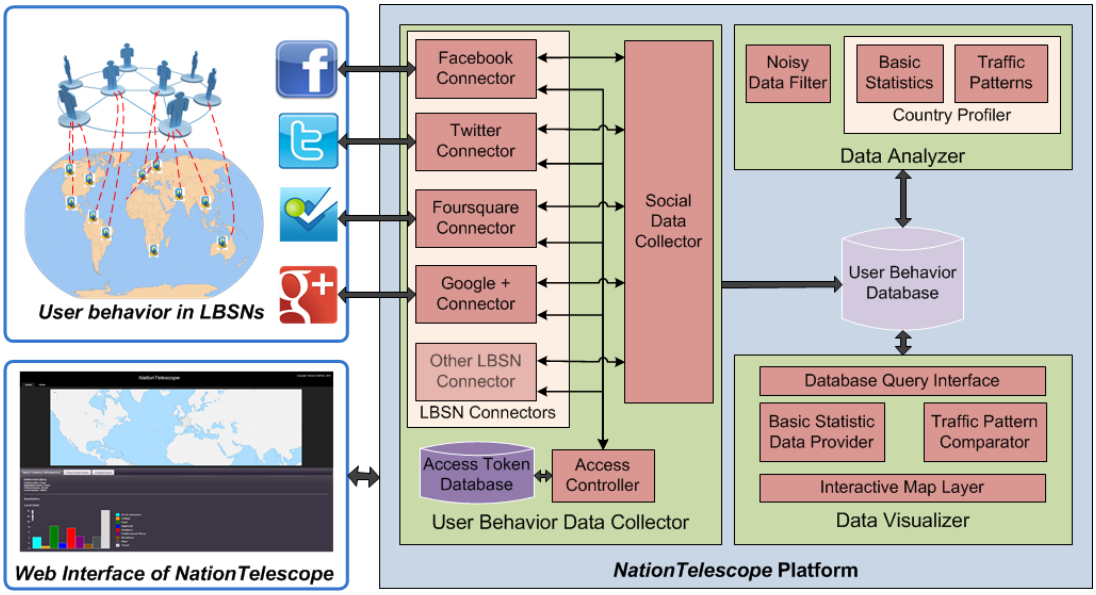


Figure 6.2 – Architecture of NationTelescope platform.

6.2 Platform Design

In this section, we present the architecture of NationTelescope platform. As shown in Figure 6.2, it mainly consists of four parts, viz., User Behavior Data Collector, Data Analyzer and Data Visualizer, as well as a User Behavior Database. First, when users interact with LBSNs, they voluntarily report their behavior data online. This data is then collected by the User Behavior Data Collector and stored in the User Behavior Database. Second, the Data Analyzer regularly accesses the User Behavior Database to conduct basic analysis, and generates summarization of the collected behavior data, such as POI visiting patterns. The summarized data is also stored in the User Behavior Database. Third, the Data Visualizer provides various visualizations of the summarized collective behavior data. In the following, we present the design and characteristics of each part.

6.2.1 User Behavior Data Collector

The User Behavior Data Collector is responsible for collecting user behavior data (i.e., check-in data) from various LBSNs services. As illustrated in Figure 6.2, it is composed of several LBSN Connectors, a Social Data collector, an Access Controller and an Access Token Database. It is borrowed from the data collection platform which is presented in Chapter 3. Please refer to Chapter 3 for more details.

6.2.2 Data Analyzer

The Data Analyzer component is responsible for conducting some basic data analysis tasks including noisy data filtering and country profiling in terms of nation-wide collective behavior. As shown in Figure 6.2, it is composed of Noisy Data Filter, and Country Profiler.

First, the raw check-in data stream from LBSNs usually contains various types of noisy data which need to be eliminated by the Noisy Data Filter. For example, some check-ins are conducted without the POI semantic information (e.g., POI category and description). Since the semantic information is indispensable for understanding collective behavior, these check-ins are thus considered as noise and need to be filtered out.

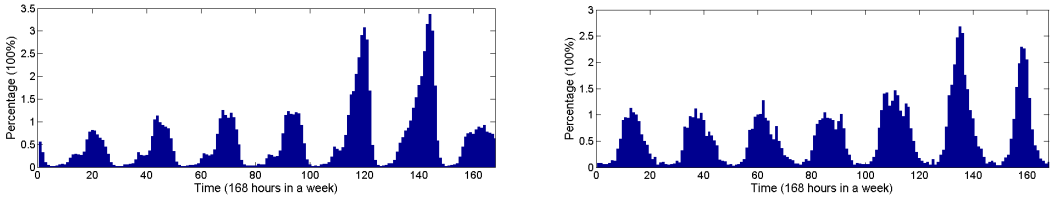
Second, based on the filtered check-in data, the Country Profiler component extracts various features to characterize the collective behavior in each country. Specifically, two types of features are extracted, viz., basic statistics and traffic patterns. The Basic Statistics of a country include the total number of check-ins, and the check-in frequency of each category of POIs (i.e., percentage of check-ins in individual POI categories). In LBSNs, POIs are classified into different categories. For example, Foursquare organizes its POIs with a three-level hierarchical category classification. It contains 9 root categories which are further classified into 291 categories at the second level. However, only a part of second-level categories are divided into sub-categories at the third level. Due to the incompleteness of the third-level categories, we choose to use the first-level and second-level categories to semantically characterize collective behavior in each country and calculate the visiting frequency for each category in a country.

In addition, in order to capture the temporal aspect of collective behavior, we also extract the traffic patterns in a "typical week" [2] for each POI category in a country, which are represented by the percentages of check-ins in each hours in a week. Figure 6.3 presents the traffic patterns of two categories of POIs, i.e., bars and museums, in the United States. We observe clearly their differences: bars are frequently visited in the evening and their traffic peaks appear on Friday and Saturday evening, while museums are often visited during daytime and the traffic peaks appear on the weekend.

6.2.3 Data Visualizer

The Data Visualizer is responsible for providing user behavior data visualization in an interactive manner, i.e., via an interactive map. As presented in Figure 6.2, it is composed of four components. First, the Database Query Interface provides the access to the User Behavior Database. Second, the Basic Statistic Data Provider takes charge of fetching the basic statistics from a country profile and normalizing the data for visualization. Third,

2. We extract weekly mean traffic patterns with hour granularity.



(a) Bar (b) Museum

Figure 6.3 – Examples of traffic patterns in the United States.

for each POI category, the Traffic Pattern Comparator provides the detailed comparison of traffic patterns from different countries. In order to efficiently identify the significant behavioral differences between countries, it incorporates a sliding-window based discriminative traffic pattern search method that will be elaborated in the next section. Fourth, all the data visualization are built upon the Interactive Map Layer. The interactive map is implemented using Leaflet [3], a light-weight cross-platform JavaScript library for interactive maps.

The design of the Data Visualizer ensures the scalability when adding new visualization components, such as tag cloud, bar chart and line chart, etc. Specifically, due to the fact that the data access and basic visualization are ensured by the Database Query Interface and Interactive Map Layer respectively, the new visualization components can be easily developed using the above two interfaces.

6.3 Platform Functionalities

In this section, we present the main functionalities of NationTelescope platform and the associated graphic user interface. Specifically, we first show the basic visualization of user behavior, including the global check-in distribution on the 3D world map, the bar charts of check-in frequency of different POI categories and the tag clouds of the checked POI categories in a specific country. We then present the visualization of the traffic pattern comparison between countries and introduce the proposed sliding-window based discriminative traffic pattern search method.

6.3.1 Basic Visualization

NationTelescope platform provides basic visualization of the summarized data. First, in order to quantitatively illustrate the collective behavior in LBSNs across the world, we present the global check-in distribution on the 3D world map by leveraging the WebGL

3. `http://leafletjs.com/`

technology [4]. WebGL (Web Graphics Library) is a powerful JavaScript API for creating interactive 3D graphics and 2D graphics within web browser without the use of plug-ins. Figure 6.4(a) shows the screenshot of the global check-in distribution. The 3D earth can be rotated or zoomed. The height of the bar indicates the total check-in count. We observe that most of the check-ins happened in big cities. Furthermore, cities in Turkey, South Asia, and South America contain a large number of active LBSN users and thus show high check-in number. Similar results have also been found in an empirical study of Foursquare usage [5].

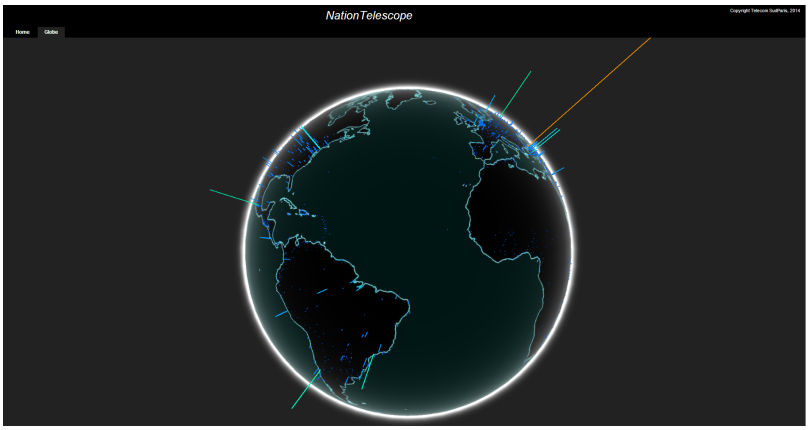
Second, we present the bar charts of check-in frequency of different POI categories. As shown in Figure 6.4(b), users can select a country by directly clicking on the interactive map. The check-in frequency of the top level POI categories is displayed as a bar chart. The bar chart is plotted using the D3.js technology [6], which is a data-driven documents JavaScript library for manipulating documents based on data. Figure 6.4(b) illustrates the check-in frequency of nine POI categories in France. We observe that travel, food, outdoor, art and entertainment related spots are frequently visited by French LBSN users.

Third, in order to understand the detailed semantics of collective behavior in a country, we look into the check-in frequency of the second-level POI categories. Due to the large number of categories (i.e., 291 categories in total), the bar chart visualization is not suitable. Therefore, we leverage the tag cloud representation to visualize the data as demonstrated in Figure 6.4(c). Larger font size of POI categories implies higher visiting frequency, and vice versa. We observe that, in Figure 6.4(c), besides the daily routine POIs, such as train stations, offices and home, French restaurants are preferred by French LBSN users.

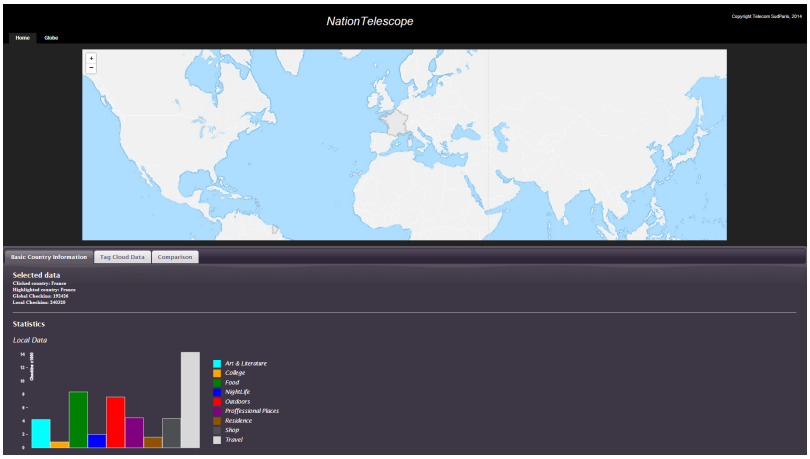
6.3.2 Traffic Pattern Visualization

In order to explore the behavioral differences between countries, NationTelescope platform supports the traffic pattern visualization functionality, which compares the traffic pattern of each POI category between two countries. However, due to a large number of POI categories (i.e., 291 categories), it is inefficient to visualize all the traffic patterns in a long list and let users explore the list to find the discriminative POI categories by scrolling the screen. Moreover, traffic patterns of some POI categories may be quite similar to each other. For example, the museum visiting patterns are probably similar in different countries. Intuitively, when comparing collective behavior between two countries, users may probably be interested in the POI categories whose traffic patterns exhibit significant difference. In the following, we first present the traffic pattern comparison scheme and then

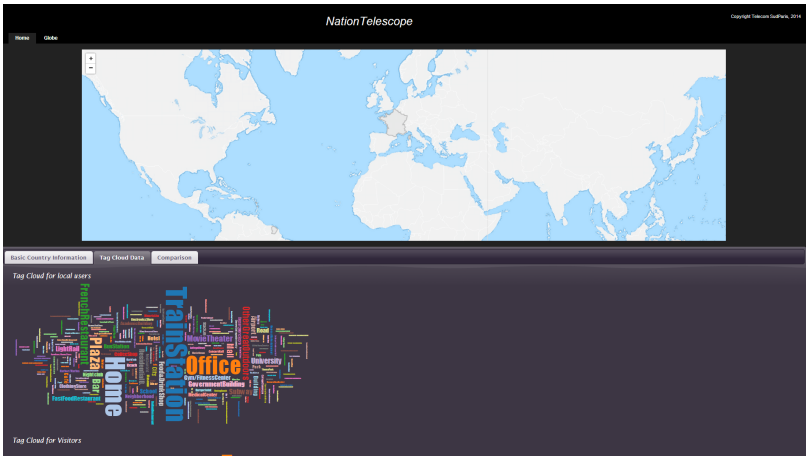
4. `http://www.khronos.org/webgl/`
5. `http://www.appappeal.com/maps/foursquare`
6. `http://d3js.org/`



(a) Global Check-in Distribution



(b) Bar charts of check-in frequency of top-level POI categories



(c) Tag Cloud of check-in frequency of second-level POI categories

Figure 6.4 – Basic visualization in NationTelescope.

demonstrate the graphic user interface for traffic pattern visualization.

6.3.2.1 Traffic Pattern Comparison

For two given countries, in order to identify the most discriminative POI categories whose traffic patterns are significantly different, we propose a sliding-window based discriminative traffic pattern search method. Specifically, for each POI category in two countries, we first normalize the two traffic patterns and then use a sliding-window to compare them in order to identify the discriminative traffic patterns and the associated difference measures. Finally, we calculate the overall difference between the two traffic patterns by averaging the difference measures of the detected discriminative traffic patterns. Figure 6.5 shows the detailed traffic pattern comparison scheme with an example of "Concert Hall" category between the United States and Japan. We present the details of each step as follows.

First, in order to focus on the temporal regularity of traffic patterns, we need to normalize the raw traffic patterns. Specifically, the raw check-in traffic patterns in countries are influenced by the number and the activeness of the users, which cannot be directly compared. For example, it is inappropriate to compare the raw traffic patterns between a country with a large number of users and that with a small number of users. Therefore, in order to avoid the influence of the number of users and the activeness of them, we normalize each traffic pattern with regard to its total number of check-ins.

Second, given two normalized check-in traffic patterns, in order to quantitatively measure the differences between them, we detect discriminative traffic patterns using a sliding-window based discriminative traffic pattern search method. Similar idea of discriminative feature selection has been widely used in various data mining problem, such as classification [32]. Specifically, we first leverage a sliding-window to compare the traffic patterns segment-by-segment to calculate their distance in each segment, and then detect the discriminative traffic patterns where the peaks of distance in all segments appear. In this work, we empirically set the size of the sliding-window as 6 hours and use Euclidean distance to quantitatively measure the difference in a segment between two traffic patterns. For example, as shown in Figure 6.5, we observe that the discriminative patterns appear in every evening. By investigating the normalized check-in traffic patterns, we see that Japanese usually go to concert earlier than Americans do in the evening.

Finally, we calculate the average difference of all discriminative traffic patterns and regard it as the overall difference between two countries with regard to a specific POI category. By calculating the difference for all the POI categories, we are able to assess how discriminative the individual POI categories are. In this work, we consider the top k most discriminative POI categories to display in the user interface, which are presented in the

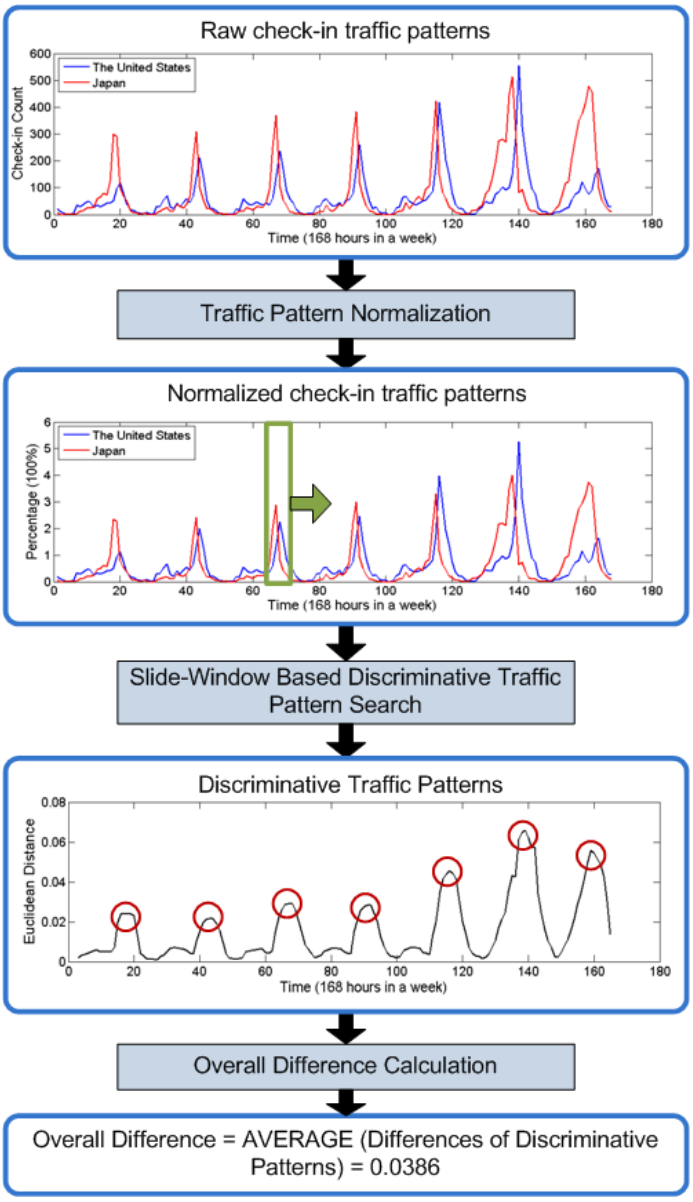


Figure 6.5 – Traffic pattern comparison scheme and an example of "Concert Hall" category comparison between the United States and Japan.

Figure 6.6 – Graphical user interface for traffic pattern visualization.

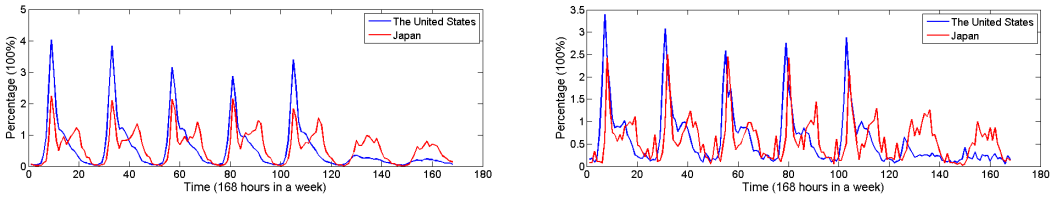
next section.

6.3.2.2 Graphic User Interface

Figure 6.6 demonstrates the Web interface for traffic pattern visualization. Users can either input the complete country names or selecting them on the map. The comparison results are then visualized by different POI categories. As shown in Figure 6.6, the visualization leverages an accordion interface with 9 top-level POI categories. When selecting a top-level category, the detailed traffic patterns of the discriminative second-level POI categories, which are identified in the previous step, are illustrated. In the screenshot, we display the top five discriminative POI categories when comparing the traffic patterns between Japan and the U.S. We observe that the "Concert Hall" category appears as the 4th discriminative POIs. Specifically, the answer of the question in the introduction section, i.e., "Do Japanese people usually go to concert earlier than Americans do?", can be summarized as follows. Japanese usually go to concert in the early evening while Americans prefer to go to concert in the late evening. In addition, more Japanese go to concert on Sunday than Americans usually do.

6.4 Evaluation

In this section, we evaluate the effectiveness and usability of NationTelescope platform. Specifically, in order to accumulate representative user behavior data for evaluation, we first implement the prototype of the platform and keep it running for about 6 months



(a) Office (b) Factory

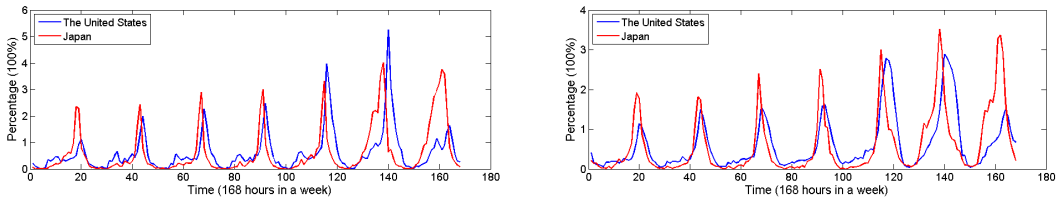
Figure 6.7 – Traffic pattern comparison of working behavior between the United States and Japan.

(from January to June 2014). We then evaluate it from the effectiveness and usability perspectives. First, in order to validate the effectiveness of NationTelescope platform, rather than exhaustively presenting behavioral comparison across all countries, we conduct two case studies and present some interesting observations. The first case study compares collective behavior between an occidental country, i.e., the United States, and an oriental country, i.e., Japan, while the second case study compares two European countries, i.e., the United Kingdom and France. Second, in order to evaluate the usability of the platform from user experience perspective, we carry out a System Usability Scale (SUS) survey with 18 participants. In the following, we first present the case studies and then the SUS study.

6.4.1 Case Study I: The United States and Japan

According to social science study, the geographical isolation is an important factor for cultural diversity [7], which leads to the behavioral difference [56]. Therefore, we choose the United States and Japan in this case study since they are geographically distant. By exploring the behavioral differences between the United States and Japan using our platform, we discover a lot of behavioral differences across various daily activities, such as working, entertainment, eating and shopping, etc. Instead of exhaustively listing all the differences, in the following, we present some interesting findings from working and entertainment behavior perspectives.

First, we find that Japanese work longer than Americans in general. We demonstrate in Figure 6.7 the traffic patterns of two POI categories (i.e., office and factory) among the top five discriminative working-related POI categories. We observe that Japanese daily working time is obviously longer than that of Americans, and a large number of Japanese work particularly in the evening. In addition, there are a lot of Japanese users working during the weekend. Similar observations have also been found in social and economy science with respect to Japanese working time [18] and the comparison of working time across different countries [64]. Compared to these traditional approaches in social and economy science



(a) Concert Hall (b) Music Venue

Figure 6.8 – Traffic pattern comparison of entertainment behavior between the United States and Japan.

that mainly consist of a large-scale survey, the advantages of NationTelescope are that it can provide timely results with significant less human effort.

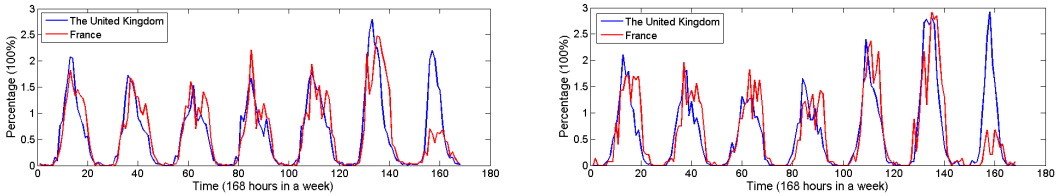
Second, we find that Americans usually go to entertainment places later than Japanese in the evening, and less Americans prefer Sunday for entertainment activities than Japanese do. As shown in Figure 6.8, we illustrate the traffic patterns of two POI categories (i.e., concert hall and music venue) among the top five discriminative entertainment related POI categories. We observe that the traffic peaks in the United States appear later than those in Japan. In addition, there is significant less traffic of entertainment activities on Sunday in the United States than that in Japan.

6.4.2 Case Study II: The United Kingdom and France

Although geographical isolation of two countries usually implies behavioral differences between the two populations, the collective behavior in two adjacent countries may still be different in some aspects. Therefore, in this case study, we choose the United Kingdom (UK) and France as the subjects of inquiry, and use our platform to explore the behavioral differences between them. In the following, we present notable differences from shopping and nightlife aspects.

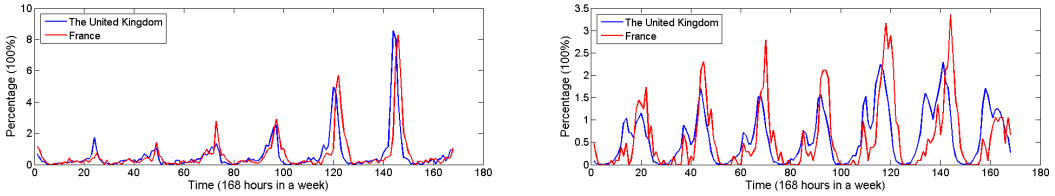
First, we find that shopping activities on Sunday are significantly less in France than those in the UK. Specifically, as shown in Figure 6.9, we demonstrate two shopping related POI categories identified by our platform as the discriminative ones, i.e., “Mall” and “Department Store”. We observe that both categories have much lower traffic in France than that in the UK. This is mainly caused by the different “rules of games” (i.e., laws) in the two countries. Nowadays, the United Kingdom opens its shops on Sunday, while France have managed to keep most of theirs closed [122].

Second, we find that the French nightlife activities are generally later than those in the UK. As shown in Figure 6.10, we present two POI categories from the top five discriminative nightlife related POI categories, i.e., nightclubs and pubs. We observe that the traffic peaks



(a) Mall (b) Department Store

Figure 6.9 – Traffic pattern comparison of shopping behavior between the United Kingdom and France.



(a) Nightclub (b) Pub

Figure 6.10 – Traffic pattern comparison of nightlife behavior between the United Kingdom and France.

in nightclubs and pubs in France appear later than the peaks in the UK.

6.4.3 Usability Study

In this study, we adopted the SUS developed by Brooke [22] in 1996, which had been widely adopted by both academia and industry. It contains a ten-item questionnaire based on Likert Scale [82], where a statement is made and respondents are supposed to indicate the degree of agreement with the statement. The SUS consists of ten statements, of which odd-numbered statements are worded positively and even-numbered statements are worded negatively. To use the SUS, participants should indicate their agreement with each statement using a five-point scale from 1 (anchored with "Strongly disagree") to 5 (anchored with "Strongly agree"). Afterwards, each statement's score contribution is determined, which ranges from 0 to 4. Concretely, for positively-worded statements (1, 3, 5, 7 and 9), the score contribution is the scale position minus 1. For negatively-worded statements (2, 4, 6, 8 and 10), it is 5 minus the scale position. Therefore, higher score for positively-worded statements implies more agreement on the statements, while higher score for negatively-worded statements implies less agreement on the statements. Finally, SUS yields a single score representing the overall usability, which is calculated by multiplying the sum of the statement score contributions by 2.5. Thus, the overall SUS score is range

Table 6.1 – System Usability Scale Scores (Higher scores imply better user experience. Note that the SUS scores for S1-S10, learnability and usability range from 0 to 4, while the overall SUS score ranged from 0 to 100.)

SUS Statements	Average Score
S1: I think that I would like to use this system frequently.	2.20
S2: I found the system unnecessarily complex.	2.87
S3: I thought the system was easy to use.	3.20
S4: I think that I would need the support of a technical person to be able to use this system.	3.13
S5: I found the various functions in this system were well integrated.	2.20
S6: I thought there was too much inconsistency in this system.	2.93
S7: I would imagine that most people would learn to use this system very quickly.	2.93
S8: I found the system very cumbersome to use.	3.06
S9: I felt very confident using the system.	2.80
S10: I need to learn a lot of things before I could get going with this system.	3.20
Learnability dimension (S4 and S7)	3.03
Usability dimension (other 8 statements)	2.81
Overall SUS score	**71.33**

from 0 to 100. Higher scores imply better user experience.

In addition, Lewis et al. [79] conducted factor analysis on the SUS statement and then defined two dimensions, i.e., learnability and usability. According to their analysis, the learnability dimension includes the statement 4 and 10 while the usability dimension includes the statements 1, 2, 3, 5, 6, 7, 8, and 9. Please refer to [22] and [79] for more details.

We conducted a SUS survey of NationTelescope platform using Google Forms [7], and spread the survey via email and social network. Participants are provided with a brief guide of the platform functionalities and are required to use the platform for about 15 minutes before they start the survey. We also provided participants with some example tasks to let them better explore our platform, such as "finding the entertainment behavioral differences between the United States and Japan". In addition, in order to collect rich user feedback, we also allow participants to leave their comments about the platform in text. We recruited 18 participants in total, of which five were female. Most of the participants (i.e., 15 participants) are between 20-30 years. The professions of the participants are diverse, including computer scientists, engineers, university students, marketing managers, etc.

Table 6.1 shows the SUS statements and the results. Higher scores imply better user experience. We find that the overall SUS score is 71.33. According to the study of Bangor et al. [10] on adjective ratings (i.e, worst imaginable, awful, poor, OK, good, excellent, best imaginable) and SUS scores (from 0 to 100), our NationTelescope platform achieves

7. `https://docs.google.com/forms`

a "good" SUS rating. Furthermore, our platform achieves high score for both usability and learnability. Specifically, the usability score and learnability score are 3.03 and 2.81 respectively. Note that the scores are ranging from 0 to 4, and higher scores imply better user experience.

In addition, by investigating the results of individual questions, we find that S3, S4, S8 and S10 have high scores, while S1 and S5 have relatively low scores. On the one hand, we understand that the NationTelescope is user-friendly and easy to use without specific preliminary knowledge. On the other hand, we understand that not all the participants would like to use the platform frequently and think that the integration of the platform can be further improved. By interviewing several participants, we find that the low score of S1 is mainly due to the fact that some of the participants are not interested in social network services, nor using social media. Moreover, some participants also mentioned that they would use the platform frequently if the platform can be integrated as an application in existing social network services, such as Facebook. According to users' comments, we find that the low score of S5 is also because they expected more integration of our platform with the existing social networks. Although it is not one of our original objectives in developing NationTelescope platform, we still plan to extend the current prototype as an integrated application in existing social networks in the future.

6.5 Discussion

Data bias in LBSNs. While the evaluation shows that NationTelescope platform can efficiently monitor and visualize large-scale collective behavior, we are aware that the platform has several limitations with respect to data bias. First, since users voluntarily report their activities in LBSNs, a user's check-in data is sparse and may not necessarily reflect her complete activity traces. Instead of considering individual behavior data, we study collective behavior data by country and show that the collective behavior is still representative and valuable. Second, the collective behavior in LBSNs may be biased due to the targeted population. Concretely, users of LBSNs mainly consist of youngsters who frequently use social network services. Therefore, the collective behavior of such a population in a country may not be completely representative of the collective behavior of its whole population. However, it can still reflect the nation-wide collective behavior to some extent. Based on our case studies, we still find some interesting behavioral differences using NationTelescope.

Seasonality of collective behavior. Intuitively, collective behavior usually exhibits seasonality [89]. For example, due to the late sunset in summer, people may probably start nightlife activities later than they do in winter. Figure 6.11 demonstrates the comparison

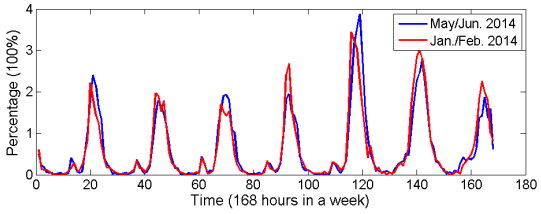


Figure 6.11 – An example of collective behavioral seasonality (Traffic pattern comparison of bars in Japan in summer (May/Jun. 2014) and that in winter (Jan./Feb. 2014)).

between the traffic patterns of bars in Japan in summer (May/Jun. 2014) and that in winter (Jan./Feb. 2014). We observe that the traffic peaks appear slightly later in summer than in winter. By studying such seasonality of collective behavior, we can conduct better behavioral comparison between countries. Therefore, although we do not fully explore such characteristics in this study due to the limited data collection time (i.e., about half a year), we plan to explore more about the seasonality of collective behavior in LBSNs and integrate new features in NationTelescope in the future.

6.6 Concluding Remarks

In this chapter, we introduce NationTelescope, a platform that monitors, compares and visualizes large-scale nation-wide user behavior in LBSNs. First, it collects the user behavior data in the check-in streaming from LBSNs. Second, it automatically generates the behavior data summary and integrates an interactive map interface for visualization. Third, it supports the collective behavior comparison functionality that detects and visualizes the discriminative behavioral differences between countries. To evaluate the effectiveness and usability of NationTelescope, we conduct two case studies and a system usability scale survey. The case studies show that our platform can efficiently capture and visualize the nation-wide collective behavior in LBSNs. The SUS survey with 18 participants proves that the our platform achieves good usability.

In the future, we plan to extend NationTelescope platform in several directions. First, according to the SUS survey results and the participants' comments, we plan to better integrate NationTelescope with the existing social network services in order to improve the user experience. Second, since NationTelescope platform continuously collects user behavior data, we intend to study the behavioral seasonality and evolution over time in the future. Third, as collective behavior in a country massively reflects the cultural information of its population, we will explore more about the correlation between global cultures and collective behaviors in LBSNs, which are presented in the next chapter.

The work in this chapter has not been previously published.

Chapter 7

Discovering Global Cultures from City-wide Collective Behavior

Contents

7.1 Introduction

Culture plays an important role in human evolution. It shapes both people belief system and practical behavior, which further solidify and evolve the culture. In the long history

of human development, there have been literally thousands of cultures on Earth, which differ in various aspects such as moral values, religious beliefs, language, clothing, cuisine, recreation, architecture, music and dance, etc. When studying culture, one of the primary tasks is to understand cultural difference across the world, which is valuable in many fields and can then support various applications. For example, knowing cultural difference between countries may help multi-national corporations explore new markets abroad.

7.1.1 Cultural Mapping and Collective Behavior

In order to identify and analyze cultural difference across the world, the United Nations Educational, Scientific and Cultural Organization (UNESCO) uses *Cultural Mapping* [112] as a crucial tool and technique to visualize cultural differences and boundaries on the map. Basically, the goal of *Cultural Mapping* is to create the map representation of different cultures and their boundaries, from the perspectives of *indigenous and local people* with respect to various cultural aspects. For example, Heatwole [55] created a world cultural map based on the people's religious beliefs. Inglehart et al. [58] built a cultural map of 53 societies based on the people's moral values extracted from the World Values Survey [1] (WVS). In order to collect the cultural related data, traditional cultural mapping approaches encompass a wide range of activities in data collection, which is mainly achieved via large-scale surveys (i.e., questionnaires and interviews) about the moral values and beliefs of the participants. For example, one question in WVS asks participants to rate the importance of family, friends, leisure time, politics, work and religion in their daily lives.

However, cultural data collection via large-scale surveys usually incurs a significant cost of both human resources and time. For example, the current wave of the WVS was carried out from 2010 to 2014, involving more than 60 countries with over 1000 participants from each country. Therefore it is hard to keep the generated cultural map up-to-date. Moreover, while such a survey is able to reveal the cultural difference from the human belief perspective, it is hard to reveal cultural difference from the human behavioral perspective.

In current literature, various definitions of culture are tightly associated with human behavior. For example, Hoebel [56] describes culture as an integrated system of learned behavior patterns which are characteristic of the members of a society and which are not a result of biological inheritance; McGrew [92] considers culture to be group-specific behavior that is acquired, at least in part, from social influences; Taylor [135] defines culture as a mental phenomenon, consisting of the contents of minds, not of material objects or observable behavior. Despite the difference in these definitions, we understand that human behavior (particularly collective behavior [138]) and culture are mutually influenced by each other. In other words, it is human behavior which solidified over generations to become that

1. `http://www.worldvaluessurvey.org`

population's culture and that culture influences further generations, and defines the human behavior of that population. The interplay of human behavior and culture motivates us to explore the cultural difference from the human behavior perspective.

Furthermore, since culture is usually studied with regard to a group of people (e.g., the whole population in a country, a state or a city), firstly we need to define an appropriate granularity for culture study, i.e., the group of people considered to have the same culture. The existing cultural mapping works are mainly conducted with country granularity, i.e., distinguishing cultures across different countries, such as that in WVS. However, in practice, culture is usually beyond country borders. On the one hand, culture often spreads across the country borders due to various cultural exchange activities, such as immigration [78], trade [48] or missions [23]. On the other hand, culture also shows diversity in the countries with large territories due to geographical isolation [7]. Therefore, in this work, we focus on analyzing the collective behavior with city granularity. For example, by analyzing the collective eating behavior in cities, we can understand the different culinary cultures across the world; by studying the statistics of spoken languages in cities, we can discover their linguistic differences. However, in practical terms, it is difficult to monitor large-scale collective behavior.

7.1.2 Cultural Features of Collective Behavior in LBSNs

With the soaring popularity of Location Based Social Networks (LBSNs), large-scale user behavioral data becomes attainable. As a typical participatory sensing system where individuals use mobile devices to share sensed data [24], LBSNs provide users with opportunities to share their real time presence with their friends by checking in at a Point of Interest (POI), such as a French restaurant or a bar, along with a short check-in message associated with their current status. By interacting with LBSNs, users generate a significant volume of check-in data online. Such large-scale user behavioral data massively reflects the cultural difference between cities across the world.

— First, check-ins at POIs in a city imply the daily activity pattern in that city. Specifically, in LBSNs, a POI is usually associated with a category, which can be considered as the semantic representation of user activities. For example, checking in at a French restaurant often means that the user is having French food there. By analyzing the collective check-ins in cities, we can reveal the cultural difference between cities with regard to their daily activity pattern. For example, there are obvious differences between western cuisine and oriental cuisine [40], which leads to the difference between users' eating behavior in Paris and in Hong Kong, i.e., local users in Paris may frequently go to French or Italian restaurants while local users in Hong Kong may frequently go to Chinese restaurants or Sushi bars.

— Second, check-ins capture inter-city user mobility patterns, which can reflect the cultural similarity/difference between cities. Specifically, by analyzing the user check-ins on a global scale, we can discover user mobility between cities. Since humans are the primary carrier of culture, human mobility and migration, which are basic cultural exchange activities, are fundamental ways of cultural diffusion [110]. Therefore, user mobility between cities can reflect the cultural similarity/difference between cities. Intuitively, cities with more similar cultures probably have more communication among them (i.e., a larger number of users travelling among the cities), and vice versa.

— Third, check-in messages in cities, which contain explicitly user status, imply linguistic characteristics of the cities. Concretely, as a form of human behavior, language is the principal means in human communication [124]. It expresses, embodies and symbolizes human culture and can thus reflect cultural differences [68]. In order to understand the language usage in LBSNs in a specific city, we conduct language detection of check-in messages. By comparing the language usage between cities, we may discover the cultural difference between cities with regard to the linguistic aspect.

Although check-ins in LBSNs contains rich cultural information, not all of them are eligible for cultural mapping. Concretely, cultural mapping suggests that only *indigenous and local people* are eligible to represent local culture [112]. Therefore, for a specific city, check-ins generated by non-local users, who are not representative for local culture, are considered as noisy data and should thus be removed when studying the city's culture. Different from a survey where we can delicately select the local people as participants, LBSNs do not allow us to select only local participants or their check-ins in the data collection process. Therefore, for a specific city, the collected check-ins often include non-local users' behavior which are not eligible in characterizing the city's culture and should thus be eliminated for cultural mapping.

7.1.3 Our Contribution: Participatory Cultural Mapping

In this chapter, aiming at discovering global cultures from collective behavior perspective, we propose a participatory cultural mapping approach, based on collective behavior in LBSNs. Specifically, the proposed approach consists of four steps. First, in order to collect large-scale user behavioral data, we collect check-ins in LBSNs on a global scale. Second, in order to detect the local users of a city, we propose a progressive "home" [2] location identification method which searches for a user's most frequented region and progressively narrows the region down to a small area. Third, by extracting the three key cultural features from

2. By "home" location of a user, we mean the location around where most of the user's activities happened rather than the actual home of the user.

local users' check-ins, i.e., daily activity pattern, inter-city mobility and linguistic feature, we propose a cultural clustering method that builds an affinity matrix between cities based on the extracted features and then leverages spectral clustering techniques to discover the cultural clusters. Finally, we generate a cultural map by visualizing the detected cultural clusters on the world map.

We experimentally evaluate the proposed approach based on a large-scale check-in dataset collected from Foursquare. Specifically, we first conduct qualitative analysis on the cultural maps created using individual features and their combination. We then quantitatively compare our cultural maps with those created by the traditional cultural mapping approaches based on survey data. The results show that the proposed approach can efficiently capture cultural information from user check-in data and generate representative cultural maps.

7.2 A Brief Review of Cultural Difference and Cultural Mapping

In thousands of years of human development, there have been thousands of cultures on Earth, which lead to cultural diversity around the world. On the one hand, cultural diversity can benefit human development. For example, different cultures usually imply different ways of thinking and solutions to problems, which is an important source of creativity. On the other hand, cultural diversity may also be a barrier in human development. For example, in the context of globalization and economic development, the lack of cultural understanding has often backfired, resulting in ineffective projects and wasted investments. Therefore, it is crucial to understand cultural differences across the world. In current literature, cultural differences have been widely studied from various domains, such as psychology [38], genetics [73], behavior [136], education [57], economy [44] and business management [95], etc.

In order to analyze cultural differences, UNESCO uses *Cultural Mapping* [112] to identify and visualize cultural differences on the map. In current literature, most of the existing works focus on cultural mapping from the psychological perspective and its applications. For example, Schwartz [125] investigated cultural differences from the value orientation perspective. Bond et al. [20] studied the cultural mapping based on human beliefs and its application to a social psychology involving culture. The Department of communications and the Arts in Australia [106] studied the cultural and economic development using cultural mapping. Evans et al. [46] applied culture mapping on arts facilities and activity planning in the UK. However, these works not only encompass an expensive data collection process via large-scale surveys, but also fail to consider people's practical behavior which is

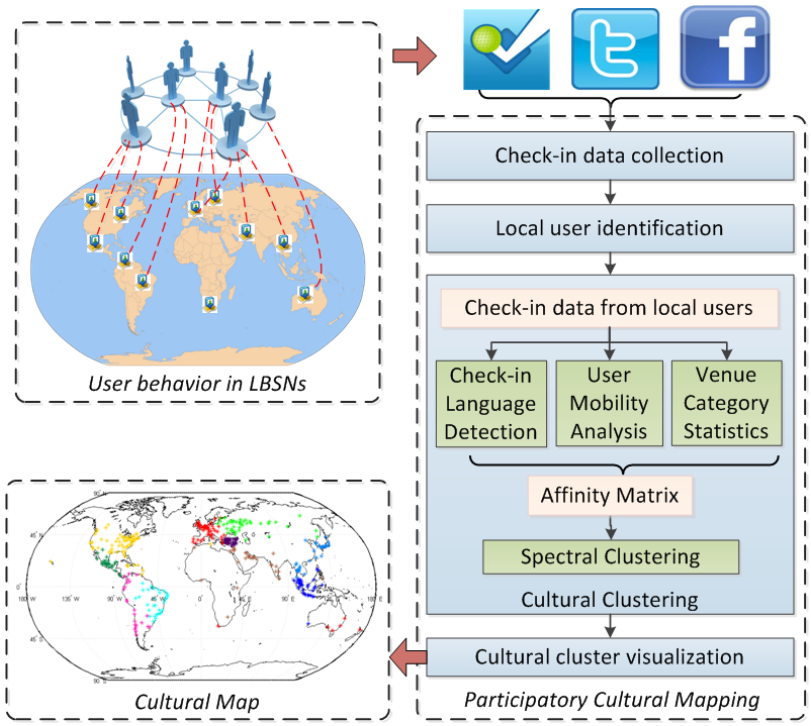


Figure 7.1 – Overview of the participatory cultural mapping approach.

also an important factor in culture [130]. In this chapter, we explore global cultures using city-wide collective behavior in LBSNs.

7.3 Overview of the Participatory Cultural Mapping Approach

Figure 7.1 illustrates the overview of the proposed participatory cultural mapping approach that consists of four parts. First, we collect check-in data from LBSNs on a global scale, which capture large-scale user behavior around the world. Second, using a progressive "home" location identification method, we identify local users in a city, whose behavior is considered to be representative in characterizing the city's culture. Third, by extracting three key features from check-in data of local users in cities, we build an affinity matrix and leverage spectral clustering techniques to discover cultural clusters of cities around the world. Finally, we plot a cultural map by simply visualizing the detected cultural clusters on the world map.

7.4 Identification of Local Users

Cultural mapping suggests that only local users in a city are eligible in order to characterize the culture of the city. In order to identify local users in a city, we need to know the home location of each user. However, due to privacy protection, such information cannot be accessed from Foursquare. Moreover, although Twitter gives users the option to register a home location for their accounts, only a limited number of users provide such information. Worse still, Twitter allows users to describe their home location without any constraint, which makes it hard to obtain a useful home location from Twitter. For example, in our dataset, some users describe their home location as "heaven and hell" or "sitting in a tin can"; some others use country names such as "Chile" or "Brazil" as their home location. Therefore, it is necessary to algorithmically identify the home location for each user.

Intuitively, we can simply search for a small area where a user checks in most frequently and regard the center of this area as her home location. However, directly searching such a small area from a mass of check-ins may overlook some user-frequented but relatively-large areas, which then leads to inappropriate home location identification. For example, considering a New York user who frequently goes to Boston for business trips and has a high check-in frequency at a few POIs there (e.g., the office of a business partner and a nearby hotel), the identified home location may probably be in Boston although the check-ins are massively around New York and its surrounding area.

In this work, rather than directly searching for a small area to identify a user's home location, we propose a progressive home location identification method. For a specific user, it starts from searching for a large region where most of the user's check-ins happen, and then repeat the search with a reduced region size within the large region, until a small region is identified. In current literature, a similar approach [33] has been used to identify user's home location in Twitter, which first segments an area into disjoint grid cells and then recursively searches for the most-checked grid cell with the decreasing cell size. However, this method may cause inaccuracy due to the segmentation process, particularly when a densely checked area is segmented into several disjoint grid cells. Therefore, instead of searching for disjoint grid cells, we iteratively search the circular regions (with a certain radius) centered by each user checked venue and select the most checked region. By repeating this step with the decreasing radius, we finally obtain a small region where the user checks in most frequently and we regard the center of this region as the user's home location. Figure 7.2 illustrates an example of the progressive home location identification method.

Formally, for a specific user u, we denote her check-ins as A_u which contains the set of the checked venues V_u. Each venue v is associated with a physical location $v.l$ (represented

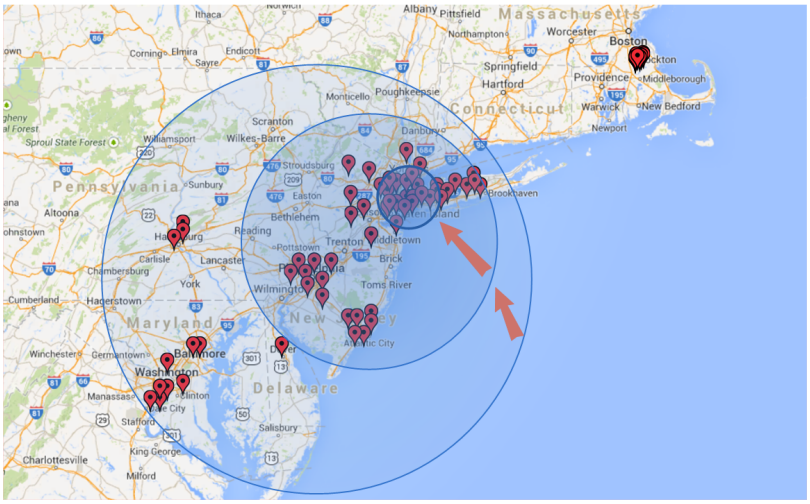


Figure 7.2 – An example of the progressive home location identification method.

Algorithm 7.5 Most-Frequented Region Search

Input: User u's check-ins A_u and search region R, target region radius d

1: Get the set of user checked venues V_u
2: Select V_u in the search region R, denoted as $V_{u,R}$
3: **for** $v \in V_{u,R}$ **do**
4: Select a region r with center $v.l$ and radius d
5: Count u's check-ins in r, denoted as $|A_{u,r}|$
6: **end for**
7: **return** $\arg\max_r |A_{u,r}|$

by GPS coordinates). In order to identify a user's home location, the proposed method aims to find a circular region $r_{l,d}$ with center l and radius d where the user has checked in most frequently. It contains a key step, i.e., the most-frequented region search, which is to search the most-frequented region $r_{l,d}$ in a given search region R. We then repeat this step with a decreasing region radius until a small region is found.

Algorithm 7.5 presents the most-frequented region search process. The basic idea is to iterate all the user checked venues in the search region R and count its surrounding check-ins in order to find the most checked region. Specifically, given a search region R and a target region radius d, we first get the venues in R where the user has ever checked in, denoted as $V_{u,R}$ (Line 1-2). Afterward, for each $v \in V_{u,R}$, we calculate the number of the user's check-ins in a candidate region r with center $v.l$ and radius d, denoted as $|A_{u,r}|$ (Line 3-6). Finally, we select the region r where $|A_{u,r}|$ is maximum (Line 7).

In order to identify a user's home location, we repeat the most-frequented region search process to recursively looking for a smaller region in the larger region identified from the

Algorithm 7.6 Progressive Home Location Identification

Input: User u's check-ins A_u, a set of target region radius D

1: Initialize the search region R as the global scale
2: Sort D in descending order
3: **for** $d \in D$ **do**
4: Search the most-frequented region r with radius d in R by Algorithm 7.5
5: Set the next search region $R = r$
6: **end for**
7: **return** The center l of r

previous iteration. Algorithm 7.6 presents the progressive home location identification method. The algorithm requires a predefined target region radius d for each iteration. We denote the set of the predefined target region radiuses as D. Given a user's check-ins A_u and a set of target region radius D, since we start from looking for the large most-frequented region on a global scale, we initialize the search region R as the global scale (Line 1), and sort target radiuses in D in descending order (Line 2). For each target region radius $d \in D$, we search the most-frequented region r with radius d in R (Line 4), and then set the next search region to the identified r (Line 5). At the end of the iteration, we return the center of the smallest most-frequented region as the user's home location (Line 7).

In this work, since we focus on identifying a user's home location at city granularity, we empirically select a set of radius $D = \{50km, 5km, 0.5km\}$, and perform home location identification. By randomly checking 500 users who have reported their home location information in Twitter, we find that there are 75% of the users (i.e., 375 users) who report valid home locations (such as GPS coordinates, specific address, or city names, etc.) that can be resolved by Google Maps [3] to get the related city information. By verifying the identified home location with the user reported city information, our method achieves an accuracy of 88.53% (i.e., 332 users' home cities are correctly identified). Compared with directly searching for the small region (i.e., $d = 0.5km$) which results in an accuracy of 72.27% and recursively searching disjoint grid cells [33] which results in an accuracy of 83.47%, the proposed progressive home location identification method achieves the best performance. More sophisticated methods (e.g., considering the text content) may be used to improve the performance. However, since it is not the main focus of this work, we use the proposed progressive home location identification method to identify the local users of a specific city.

3. https://maps.google.com

7.5 Cultural Clustering

By analyzing the collective behavior of the local users in cities, we can study the cultural differences between cities and discover cultural clusters based on these differences. Specifically, we first extract three key cultural features from check-in data, i.e., daily activity pattern, inter-city mobility and linguistic feature, in order to build an affinity matrix of cities. We then leverage spectral clustering techniques to discover cultural clusters based on the built affinity matrix.

7.5.1 Feature Extraction

User check-ins in LBSNs massively imply cultural information. First, the daily activity pattern in a city can be characterized by the categories of user checked POIs. Second, the inter-city mobility representing cultural exchange activities can be extracted from check-ins. Third, the linguistic feature characterized by the practical language usage can be obtained from check-in messages by leveraging language detection techniques.

7.5.1.1 Daily Activity Pattern

By collecting the check-ins of local users in a city, we are able to understand the citizen's daily activities. POI categories can be regarded as the semantic representation of users' activities when checking in. For example, checking in at a French restaurant probably means the user is having French food there. Therefore, we characterize a city's daily activity pattern by the check-in distribution on different POI categories.

Venues in Foursquare are organized with a three-level hierarchical category classification by the date of data collection. Specifically, it contains 9 root categories which are further classified into 291 categories at the second level. Moreover, a few second-level categories have sub-categories at the third level. Due to the incompleteness of third-level categories, only a few venues have the category information at third level. Therefore, we choose to use the second-level categories (291 categories) to semantically characterize users' behavior when checking in at POIs. Figure 7.3 demonstrates venue category tag clouds in New York and Tokyo. We first observe that offices, subway and train stations are popular check-in POIs in both cities. Most importantly, we observe clearly the cultural differences between the two cities, i.e., New York users usually check in at home, bars, gyms, outdoor places while Tokyo users often go to ramen/noodle houses, convenience stores, and Japanese restaurants.

Using the second level venue categories provided by Foursquare, we characterize a city's daily activity pattern using a 1×291 vector, representing the check-in distribution on the 291 venue categories. In order to quantitatively measure the difference on daily activity

(a) New York (b) Tokyo

Figure 7.3 – Examples of venue category tag clouds (Larger font size implies higher frequency, and vice versa.)

pattern between cities, we resort to the Jensen-Shannon divergence [83] to measure the difference between two probability distributions because it is a *symmetric* and *bounded* metric. Specifically, given two distributions P_1 and P_2, the Jensen-Shannon divergence is calculated as follows:

$$JSD(P_1||P_2) = \frac{1}{2}KLD(P_1||M) + \frac{1}{2}KLD(P_2||M) \tag{7.1}$$

where $M = \frac{1}{2}(P_1 + P_2)$ and $KLD(P||M)$ is the Kullback-Leibler divergence [69] which is calculated as:

$$KLD(P||M) = \sum_i \log_2(\frac{P(i)}{M(i)}) \cdot P(i) \tag{7.2}$$

We see that the Jensen-Shannon divergence can be regarded as the symmetrized and smoothed version of the Kullback-Leibler divergence. Using the base 2 logarithm, the Jensen-Shannon divergence is bounded in $[0, 1]$ [83]. Therefore, for two specific cities C_1 and C_2, given their check-in distribution on venue categories P_{C_1} and P_{C_2}, we define the daily activity pattern similarity between the two cities, denoted by Sim_{DAP}, as follows:

$$Sim_{DAP}(C_1, C_2) = 1 - JSD(P_{C_1}||P_{C_2}) \tag{7.3}$$

7.5.1.2 Inter-city Mobility

By quantitatively analyzing check-ins in different cities, we are able to understand the inter-city mobility, which implies the inter-city cultural similarity/difference. Intuitively, users in two cities that share similar culture will probably be easy to communicate and interact (e.g., doing business) among them, and thus probably have more travels from one to the other. Therefore, we investigate the behavior of users who have ever checked in in multiple cities. Specifically, we first find out the fraction of users in a city who have ever

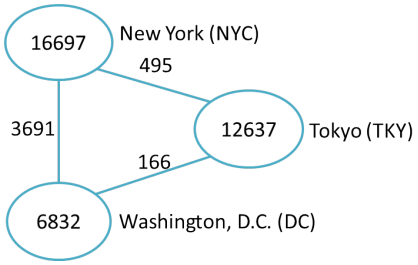


Figure 7.4 – An example of inter-city mobility between New York, Washington D.C. and Tokyo.

been checked in in another city, and then we define a symmetric metric to measure the similarity between cities based on those fractions.

Specifically, for two given cities C_1 and C_2, we denote the users who have checked in each of them as U_{C_1} and U_{C_2}, respectively. We then calculate the fraction of U_{C_1} who have ever checked in C_2, i.e., $\frac{|U_{C_1} \bigcap U_{C_2}|}{|U_{C_1}|}$, and the fraction of U_{C_2} who have ever checked in C_1. i.e., $\frac{|U_{C_1} \bigcap U_{C_2}|}{|U_{C_2}|}$. Finally, we combine them using the geometric mean and define the similarity based on inter-city mobility, denoted by Sim_{Mob}, as follows.

$$Sim_{Mob}(C_1, C_2) = \sqrt{\frac{|U_{C_1} \bigcap U_{C_2}|}{|U_{C_1}|} \cdot \frac{|U_{C_1} \bigcap U_{C_2}|}{|U_{C_2}|}} \tag{7.4}$$

The choice of geometric mean ensures that two cities are similar if and only if there is a significant fraction of users from both cities who have travelled between them. In addition, according to the definition of the Sim_{Mob}, it is easy to prove that Sim_{Mob} is bounded in $[0, 1]$.

Figure 7.4 presents an example of three cities in our dataset, i.e., New York, Washington D.C. and Tokyo. The number inside the circle presents the city's total user number. The number on the link between cities presents the number of users who have checked both of the cities. We then calculate the similarity between them based on the inter-city mobility as follows.

$$Sim_{Mob}(NYC, DC) = 0.3456 \tag{7.5}$$

$$Sim_{Mob}(NYC, TKY) = 0.0341 \tag{7.6}$$

$$Sim_{Mob}(DC, TKY) = 0.0179 \tag{7.7}$$

Due to the cultural differences between Japan and U.S., we observe that the similarity with respect to user mobility between New York and Washington D.C. is significantly higher than that between Tokyo and New York (or Washington D.C.).

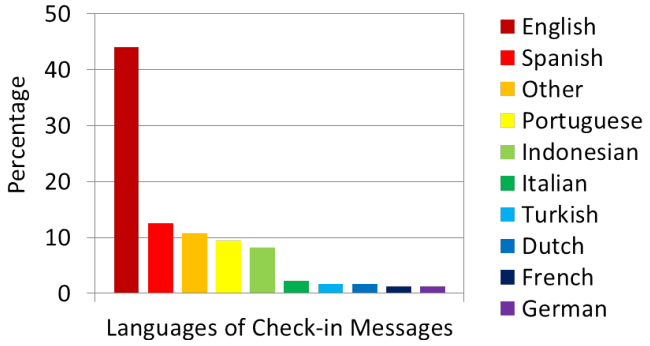


Figure 7.5 – Top 10 languages of check-in messages.

7.5.1.3 Linguistic Feature

Check-in messages massively imply the linguistic characteristics of a city, which play an important role in human culture. In this work, by conducting analysis on check-in messages, we investigate the practical language usage in LBSNs in cities. Specifically, by applying language detection techniques on check-in messages in a city, we can characterize the linguistic feature of a city by the distribution of the languages used in the city, and then quantitatively measure the difference/similarity between cities. Due to the complexity and difficulty of multilingual text analysis (e.g., multilingual sentiment analysis), which is also beyond our focus, we do not explore the content and the detailed semantically meaning of check-in messages in this work.

In order to identify the language of a check-in message, we leverage the language detection library developed by Cybozu Labs [128]. We detect 47 languages in total (including "other" for unknown languages) in our dataset. Figure 7.5 demonstrates the top 10 languages and their percentages in our dataset. Unsurprisingly, English is the dominant language used in LBSNs. Some popular languages, such as Spanish and Portuguese, also appear at the top of the list. Similar results have also been reported in [77]. By excluding the unknown languages, i.e., "other", we can characterize a city by a distribution of check-ins on 46 languages. Figure 7.6 illustrates two tag clouds of languages in two big cities, i.e., Mexico City and Rio de Janeiro. We observe that English is the most popular language in both cities in LBSNs, even though it is not the official language in either of the cities. This is due to the fact that the language in LBSNs is highly biased towards English. Specifically, we find that English is the most used language in 95% of the cities in our dataset. However, we can still discover the linguistic difference between the cities, i.e., Spanish and Portuguese are the second most popular languages in Mexico City and Rio de Janeiro, respectively.

Similar to the daily activity pattern, we leverage the Jensen-Shannon divergence to

(a) Mexico City (b) Rio de Janeiro

Figure 7.6 – Examples of language tag clouds.

measure the linguistic difference between cities. Formally, for two specific cities C_1 and C_2, we denote their distributions of check-ins on languages as L_{C_1} and L_{C_2}, respectively. We define the linguistic similarity between the two cities, denoted by Sim_{Lin}, as follows:

$$Sim_{Lin}(C_1, C_2) = 1 - JSD(L_{C_1}||L_{C_2}) \tag{7.8}$$

7.5.1.4 Affinity Matrix Construction

By characterizing the culture similarity/difference from three different aspects, i.e., daily activity pattern, inter-city mobility and linguistic feature, we combine them into a unique measure by leveraging their geometric mean.

$$Sim = \sqrt[3]{Sim_{DAP} \cdot Sim_{Mob} \cdot Sim_{Lin}} \tag{7.9}$$

It ensures that two cities are similar if and only if they are similar in all three aspects. Since all the similarity measures, i.e., Sim_{DAP}, Sim_{Mob} and Sim_{Lin}, are bounded in $[0, 1]$, it is easy to prove that Sim is also bounded in $[0, 1]$. For a given set of cities, by calculating all the similarities between each pair of cities, we can then construct an affinity matrix in order to discover cultural clusters from it.

7.5.2 Spectral Clustering

Given an affinity matrix measuring the cultural similarity between cities, we adopt the spectral clustering techniques [139], which are widely adopted in various clustering problems due to the quality of the clusters generated and the simplicity of implementation. We use a variation of spectral clustering proposed in [96] which integrates a normalization step and shows better performance compared to the classical spectral clustering algorithm [139]. Moreover, similar to [41], we also integrate a method to auto-select the number of clusters within a given range.

Algorithm 7.7 presents the clustering process. Let M denote an affinity matrix of cities, with the size of $n_c * n_c$, where n_c is the number of cities. We also define a range

Algorithm 7.7 Spectral Clustering with Auto-Selected Number of Clusters

Input: Affinity matrix M, Range of the number of clusters $[k_{min}, k_{max}]$

1: Construct the diagonal degree matrix D that $D(i,i) = \sum_{j=1}^{n_c} A(i,j)$
2: Calculate the Laplacian matrix $L = D - A$
3: Calculate the normalized Laplacian matrix $L_{norm} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$
4: Get the k_{max} smallest eigenvalues $\{\lambda_1, ..., \lambda_{k_{max}}\}$ of L_{norm}
5: Calculate $\delta_i = \lambda_{i+1} - \lambda_i$ for $i \in [k_{min}, k_{max} - 1]$
6: Select the optimal $k = \arg\max_i \delta$
7: Get the k smallest eigenvectors $\{e_1, ..., e_k\}$ of L_{norm}
8: Construct a matrix X where e_i is its ith column
9: Treat each row of X as a data sample and cluster them into k clusters using k-means, denoted as $\{C_1, ..., C_k\}$
10: **return** $\{C_1, ..., C_k\}$

for the number of clusters, denoted as $[k_{min}, k_{max}]$, as the inputs. In order to conduct spectral clustering, we start by calculating the normalized Laplacian matrix (Line 1-3). We then select the optimal number of clusters k in $[k_{min}, k_{max}]$ by searching for the largest gap between two consecutive eigenvalues (Line 4-6). Finally, by calculating the k smallest eigenvectors and use them to represent the data samples, we adopt k-means to cluster them into k clusters (Line 7-9). Please refer to [96] for more mathematical details about spectral clustering.

Once we obtain the cultural clusters, combined with the location of the cities, we create a cultural map by visualizing these clusters using different colors on the map.

7.6 Experimental Evaluation

In order to validate the proposed participatory cultural mapping approach, we carry out various experiments based on the large-scale check-in data collected from Foursquare, and conduct both in-depth qualitative and quantitative analysis on the generated cultural maps. Specifically, by selecting 415 big cities around the world, we qualitatively study the cultural map generated using the proposed approach and the implication of the individual features, and show some interesting cultural correlations between user behavior and other factors such as geography, immigration, religion, etc. Moreover, by comparing the cultural maps (or cultural clusters) created using the survey data from the WVS [58] and the GLOBE [4] (Global Leadership and Organizational Behavior Effectiveness research) project [53], we quantitatively evaluate the proposed approach and discuss its advantages and limitations.

4. http://www.tlu.ee/~sirvir/Leadership/Leadership%20Dimensions/globe_project.html

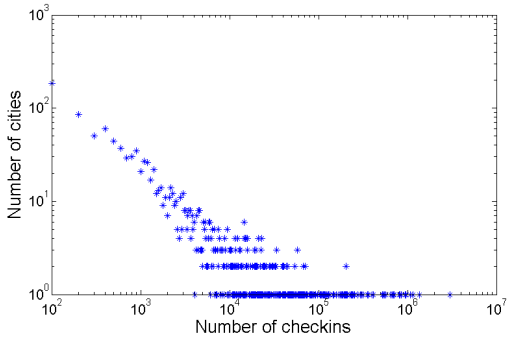


Figure 7.7 – Distribution of the number of check-ins in cities.

7.6.1 Dataset Selection

In this work, we collect check-in data in Foursquare. Using our data collection platform, we collected a check-in dataset over about 18 months (from April 2012 to September 2013). After noisy data filtering, our dataset includes 279,495 users who have performed 49,273,956 check-ins at 6,743,711 venues globally.

Since we focus on the cultural map with city granularity, we leverage a dataset of world cities provided by ESRI [5], a leading company in Geographic Information System (GIS). The dataset consists of 2535 major cities around the world, most of which are national or provincial capitals. Combined with the check-in dataset, we plot the distribution of the number of check-ins in cities in Figure 7.7. We observe that it follows a power-law distribution [37], which means that there are a large number of cities with a small number of check-ins. Intuitively, the user check-ins in these less checked cities may not be sufficient or representative enough to characterize the cities' culture. Therefore, in order to filter out the less checked cities, we select the cities containing more than 10,000 check-ins as valid cities, resulting in 415 valid cities located in 77 countries. Table 7.1 presents the statistics of the selected dataset. We observe that the 415 valid cities contain 81% of the global check-ins. Figure 7.8 illustrates the tag cloud of the country where these 415 cities are located. Unsurprisingly, the United States, where Twitter and Foursquare started their business, has most cities (i.e., 60 cities) in the dataset.

7.6.2 Qualitative Evaluation

In order to qualitatively evaluate the proposed approach, in this section, we first analyze the cultural map created with the selected dataset, and then discuss the implications of the individual features (i.e., daily activity pattern, inter-city mobility and linguistic feature) in

5. http://www.esri.com/

Table 7.1 – Dataset Statistic.

Number of valid cities	415
Number of countries	77
Number of check-ins	40,091,800
Number of users	266,909
Number of venues	3,680,126

Figure 7.8 – Tag cloud of countries where the selected cities are located.

cultural mapping. Based on the cultural maps/clusters presented in [55,58] and [53], where the number of clusters are 9, 10 and 13, respectively, we empirically set the range for the number of cultural clusters as $[5, 15]$ and let Algorithm 7.6 select the optimal one in this range.

Figure 7.9 demonstrates the cultural map created by our approach, which identifies 10 clusters in total. First, we observe the geographical constraint on cultural clusters, i.e., cities in a cluster are located in a specific region in the world. For example, some major cultural clusters are North America, Middle America, South America, Western Europe, Eastern Europe, the Middle-East, East Asia and South Asia. However, the geographical proximity between two cities does not necessarily mean that they are associated with the same cluster. For example, Cape Town in South Africa and some cities in Australia and New Zealand are associated with the Western Europe cluster, even though they are geographically distant

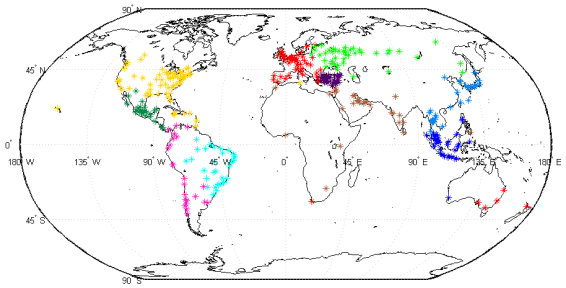
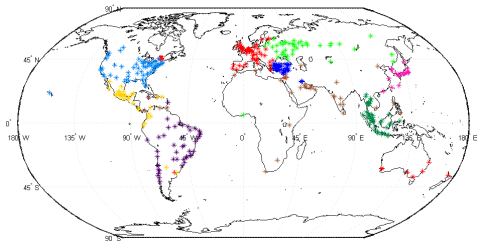
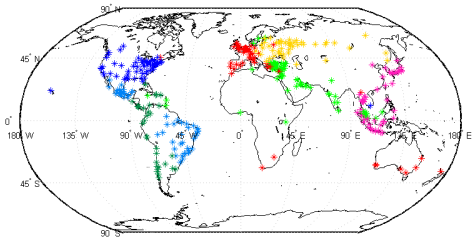


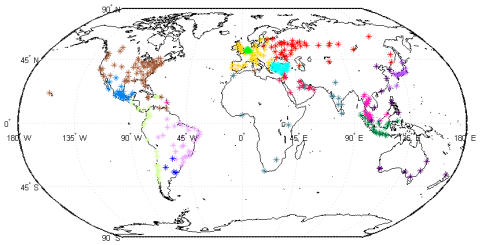
Figure 7.9 – Cultural map created by the proposed approach.



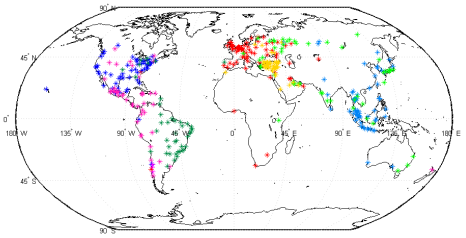
(a) Daily activity pattern



(b) Daily activity pattern with only food-related activities



(c) Inter-city mobility



(d) Linguistic feature

Figure 7.10 – Cultural maps based on individual features (Note that the colors of the clusters are assigned in the way that they can be better visually distinguished, and there is no correspondence between the clusters with the same color in different cultural maps.)

from Western Europe. It is probably due to the emigration from the UK to these cities which were its colonies in the past. Second, we observe that there are two clusters in Latin America, which are separately located in the Eastern part and Western part of Latin America. It is probably due to the language usage in Latin America, i.e., Portuguese is the dominant language in West Latin America while Spanish is the most popular language in East Latin America. Third, cities in Turkey, Greece and Cyprus form a standalone cluster surrounded by Western Europe, Eastern Europe and Middle-East clusters. By investigating the individual features of these cities, we find that there is a significant mobility among these cities, which then leads to a standalone cluster in this area.

In order to further evaluate our approach, we create cultural maps based on individual features and then study the implications of these features in cultural mapping. Figure 7.10 presents four cultural maps based on the cities' daily activity pattern, daily activity pattern with only food related activities, inter-city mobility and linguistic feature. Note that the colors of the clusters are assigned in the way that they can be better visually distinguished, and there is no correspondence between the clusters with the same color in different cultural maps.

7.6.2.1 Daily activity pattern

Figure 7.10(a) demonstrates the cultural map created based on the daily activity pattern in cities, where 9 cultural clusters are identified. Although it looks similar to the cultural map created using all the features, there are still some interesting differences. First, only one cluster dominates Latin America due to the similar activity patterns in the cities there. Second, Montreal, which was a French colony for over 200 years and is now a bilingual (i.e., French & English speaking) city in Canada, is clustered together with Western European cities. By investigating the check-in POI categories in Montreal, we find that there are a significant number of French Restaurants there, which is the primary reason that it is put in the Western Europe cluster.

Furthermore, since food is a fundamental element in a culture [40], the cultural difference of food has been studied from various perspectives, such as flavor [3] and recipe [165]. Therefore, we are motivated to study the food preferences across the world by analyzing people's eating activities reported to LBSNs. Specifically, we create a cultural map only based on the food-related activities (i.e., check-ins at the POIs of the "Food" root-category including 88 sub-categories). Figure 7.10(b) presents the cultural map based on the food-related activities. We observe that Montreal is still with the Western Europe cluster. Interestingly, there are a number of cities in South East Asia, particularly in Malaysia and Indonesia, which are in the same cluster as Middle-Eastern cities, although they are geographically distant. It can probably be explained by the religious similarity and its impact on the food preferences in those cities. While Islam is the largest religion in the Middle East, it is also the most widely practiced religion in Malaysia and Indonesia [16]. Although Middle Eastern cities are geographically distant from Malaysian and Indonesian cities, due to the same Islamic dietary law, the food preferences in these cities are similar.

7.6.2.2 Inter-city mobility

Figure 7.10(c) presents the cultural map based on the inter-city mobility, where 14 clusters are identified. We observe strong geographical constraints on the clusters. First, due to the power-law distribution of travelling distance in LBSNs [33, 100], the inter-city mobility tends to be significant within small areas, which leads to more clusters with the small geographical span. In addition, the administrative constraints (e.g., visa applications) also mean that a number of users may only travel within their own countries, which is also the reason that cultural analysis is often conducted with country granularity in current literature [53, 58].

7.6.2.3 Linguistic feature

Figure 7.10(d) presents the cultural map based on the linguistic feature, where 7 clusters are identified. We observe that the clusters do not have clear geographical boundaries between each other and thus overlap. This is due to the fact that check-in languages in LBSNs are highly biased towards English. Although the languages of check-in messages are biased representations of the languages in a city, we can still observe some interesting clusters. For example, Latin American is separated into two clusters due to the fact that Portuguese is the official language in Brazil, while Spanish is the most popular language in most of the other Latin American countries.

7.6.3 Quantitative Evaluation

Different from the traditional cultural mapping approaches that mainly collect data via large-scale surveys, we propose a participatory cultural mapping approach that leverages the participatory sensed collective behavior data. In this section, since a cultural map intrinsically consists of a set of cultural clusters, we quantitatively evaluate the proposed approach by comparing the traditional cultural clusters based on survey data and the ones generated by the proposed approach. Specifically, we first find two related works that identify cultural clusters using survey data, and then select the valid cultural clusters including the common cities in our dataset and their dataset. By applying our cultural mapping approach on the check-in data in the related cities with different features, we conduct an overall comparison between the obtained cultural clusters and the traditional cultural clusters using Normalized Mutual Information (NMI). Finally, by conducting the cluster-wise comparison, we analyze the correlation and the differences between our cultural clusters and the traditional clusters.

7.6.3.1 Traditional cultural clusters based on survey data

We have found two works related to cultural mapping using the survey data from the WVS and the GLOBE project in current literature. Specifically, based on the people's moral value data in 53 countries from WVS, Inglehart et al. [58] created a cultural map consisting of 9 cultural clusters, viz., "English Speaking", "Catholic Europe", "Protestant Europe", "Orthodox", "Latin America", "Africa", "Islamic", "South Asia" and "Confucian". Based on people's leadership psychology [13] data in 61 countries from the GLOBE Project, Gupta et al. [53] identified 10 cultural clusters, viz., "Anglo", "Latin Europe", "Nordic Europe","Germanic Europe", "Eastern Europe", "Latin America", "Arab", "Sub-Saharan Africa", "Southern Asia" and "Confucian Asia".

However, while our approach focuses on city granularity, these works all focus on the

Table 7.2 – Cultural clusters and their city numbers.

Cultural Clusters of WVS [58]		Cultural Clusters of GLOBE [53]	
Cluster Name	Number of cities	Cluster Name	Number of cities
English Speaking	83	Anglo	85
Catholic Europe	17	Latin Europe	15
Protestant Europe	21	Nordic Europe	3
Orthodox	39	Germanic Europe	19
Latin America	75	Eastern Europe	38
Islamic	40	Arab	67
South Asia	57	Southern Asia	56
Confucian	22	Confucian Asia	23
		Middle East	42

cultural clusters on country granularity. In order to bridge this gap, we consider all cities in a country to be within the same cluster and then obtain the corresponding cultural clusters on city granularity. Moreover, since our dataset contains 77 countries in total, we only use the cities that appear in both our dataset and the dataset in [58] or [53] for comparison. As a result, due to the low popularity of LBSNs in Africa, only two cities appear in the "Africa" cluster in the WVS dataset, and none of the cities appears in the "Sub-Saharan Africa" cluster in the GLOBE dataset in our dataset. Therefore, we remove these clusters and filter out the cities concerned. Finally, we obtain 8 cultural clusters with 354 cities for [58] and 9 cultural clusters with 348 cities for [53]. Table 7.2 presents the cultural clusters and the number of cities in individual clusters.

We observe that the cultural clusters in these two works have an obvious correspondence. The main difference is that Gupta et al. considered countries in "Nordic Europe" as a standalone cluster, while Inglehart et al. put them in the "Protestant Europe" cluster.

7.6.3.2 Overall comparison with traditional cultural clusters

Based on the selected cities of WVS and GLOBE Project, we identify cultural clusters using the proposed approach with individual features as well as their combination. We then calculate the Normalized Mutual Information (NMI) [4] between our cultural clusters and the traditional cultural clusters, which measures the correlation between them. Formally, for a dataset of N cities, we denote two sets of cultural clusters as $\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$ and $\Phi = \{\phi_1, \phi_2, ..., \phi_J\}$, where ω_k represents the kth cluster in Ω, and so on. The normalized mutual information is calculated as follows:

$$NMI(\Omega, \Phi) = \frac{2I(\Omega, \Phi)}{H(\Omega) + H(\Phi)} \tag{7.10}$$

Table 7.3 – Normalized mutual information between different sets of cultural clusters.

Features	WVS data [58]		GLOBE data [53]	
	Identified cluster number	NMI	Identified cluster number	NMI
Daily activity pattern	7	0.7446	7	0.7273
Inter-city mobility	8	0.7133	8	0.6796
Linguistic feature	9	0.5138	8	0.5354
All features	8	0.7669	7	0.7416

where I is the mutual information between Ω and Φ. $H(\Omega)$ and $H(\Phi)$ is the entropy of Ω and Φ, respectively. Using maximum likelihood estimation, they are calculated as follows:

$$I(\Omega, \Phi) = \sum_{k=1}^{K} \sum_{j=1}^{J} \frac{|\omega_k \cap \phi_j|}{N} \log \frac{\frac{|\omega_k \cap \phi_j|}{N}}{\frac{|\omega_k|}{N} \frac{|\phi_j|}{N}} \tag{7.11}$$

$$H(\Omega) = -\sum_{k=1}^{K} \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N}, H(\Phi) = -\sum_{j=1}^{J} \frac{|\phi_j|}{N} \log \frac{|\phi_j|}{N} \tag{7.12}$$

Note that N is the number of cities in the dataset. The value of NMI is actually bounded in $[0, 1]$. The higher value implies higher correlation between two sets of clusters. Please refer to [4] for more mathematical details.

Table 7.3 presents the experiment results. First, we observe that the numbers of the cultural clusters identified by our approach are quite similar to that of traditional cultural mapping approaches. Second, the incorporation of all features in our approach results in the best NMI and outperforms all the results using individual features. Moreover, comparing the results using individual features, we find that daily activity pattern feature results in the best NMI, followed by inter-city mobility feature. Due to the bias of language usage in LBSNs, the linguistic feature yields the worst results.

7.6.3.3 Cluster-wise comparison with traditional cultural clusters

In order to better understand the difference between our cultural clusters and traditional ones, we further analyze the correlation between each pair of clusters. Specifically, for each of our clusters, we calculate its purity with respect to each traditional cultural cluster, i.e., the percentage of its cities that appear in each of tradition cultural clusters. Table 7.4 and 7.5 present the results on the WVS and Globe dataset, respectively. C1 represents the first cluster in our cultural clusters, and so on. Taking the first column in Table 7.4 as an example, 2%, 41% and 57% of the cities in C1 belong to the "English Speaking", "South Asia" and "Confucian" cultural clusters, respectively. We highlight the highest percentage in each column, which indicates the most relevant traditional cultural cluster for each of our clusters.

Table 7.4 – Comparison of individual cultural clusters (WVS dataset).

	C1	C2	C3	C4	C5	C6	C7	C8
English Speaking	0.02	0	**0.96**	0.03	0	0	0.21	0.06
Catholic Europe	0	0	0	0	0	0	0.28	0
Protestant Europe	0	0	0	0	0	0.03	**0.34**	0
Orthodox	0	0	0	0	0	**0.97**	0.03	0
Latin America	0	**1**	0.01	**0.97**	0	0	0	0
Islamic	0	0	0	0	**0.92**	0	0.08	0
South Asia	0.41	0	0.03	0	0.08	0	0.06	**0.94**
Confucian	**0.57**	0	0	0	0	0	0	0

Table 7.5 – Comparison of individual cultural clusters (GLOBE dataset).

	C1	C2	C3	C4	C5	C6	C7
Anglo	0.03	0	**0.96**	0	0.23	0.1	0
Latin Europe	0	0	0	0	0.21	0	0
Nordic Europe	0	0	0	0	0.04	0	0
Germanic Europe	0	0	0	0	**0.28**	0	0
Eastern Europe	0	0	0	**0.86**	0.05	0	0.05
Latin America	**0.97**	0	0.01	0.14	0	0	0
Arab	0	0	0	0	0.1	0	**0.95**
Southern Asia	0	0.39	0.03	0	0.1	**0.87**	0
Confucian Asia	0	**0.61**	0	0	0	0.03	0

On the one hand, we observe that the cultural clusters identified by the proposed approach are highly correlated with the traditional cultural clusters. Specifically, some traditional cultural clusters can be obviously identified in both WVS and Globe datasets, i.e., the cities of those clusters mostly appear in only one of our clusters. For example, with the WVS dataset, "English Speaking", "Orthodox", "Islamic", "South Asia" and "Confucian" clusters are clearly associated with C3, C6, C5, C8 and C1, respectively. With the Globe dataset, "Anglo", "Eastern Europe", "Latin America", "Arab", "Southern Asia" and "Confucian Asia" clusters are clearly associated with C3, C4, C1, C7, C6 and C2, respectively.

On the other hand, we also observe some interesting differences between our clusters and the traditional cultural clusters. First, with the WVS dataset, two of our clusters, i.e., C2 and C4, are associated with "Latin America" clusters. This is mainly due to the consideration of linguistic feature in our approach, i.e., Spanish and Portuguese are two major languages in Latin America. Second, with the WVS dataset, the cities in Western Europe are put together in one cluster, i.e., 21%, 28% and 34% of the cities in cluster C7 are associated with "English Speaking", "Catholic Europe" and "Protestant Europe" clusters, respectively. A similar observation can also be found in the cluster C5 using the Globe dataset. By investigating the similarity between the related cities based on the individual

features, we found that these cities are very similar to each other with respect to the inter-city mobility and the daily activity pattern. Therefore, from the user behavior perspective, our approach puts them in the same cluster.

7.7 Discussion

Data bias in LBSNs. User check-ins in LBSN may not be a full representation of users' daily activities. Since users voluntarily report their activities in LBSNs, check-ins are biased samples of user daily activities, which can be regarded as a social representation of user activities. Moreover, the user community of LBSNs may be biased towards young people who prefer to use social network services. However, despite the existence of these data bias in LBSNs, our study shows that check-in data still contains valuable cultural information and can be used to generate representative cultural map. In the future, we plan to investigate more into the influence of these data bias on the cultural mapping.

Temporal dynamics of culture and collective behavior. It is known that culture can spread from one area to another due to the various cultural exchange activities, such as immigration. In the human history, such culture diffusion process is usually slow over time and gradually leads to globalization [119]. In this study, due to the limited duration of user behavioral data collection process, we do not investigate the temporal dynamics of culture and collective behavior. However, as we continuously collect social media data, we believe that in the future, we can track cultural diffusion by studying long-term user activity data in LBSNs.

7.8 Concluding Remarks

Cultural mapping has been recognized as a crucial tool by UNESCO to visualize cultural difference and culture boundaries on the map. Traditional cultural mapping approaches usually rely on large-scale survey data with respect to human belief, which fall short due to the expensive data collection process and lack of capturing human behavior. In this chapter, aiming at studying the correlation between collective behavior and human cultures, we propose a participatory cultural mapping approach based on the collective behavior in LBSNs. Specifically, we first collect user participatory sensed behavioral data from LBSNs and then filter out noisy data from non-local users. Afterwards, by collecting the three key features, i.e., daily activity pattern, inter-city mobility and linguistic feature, we propose a cultural clustering method based on spectral clustering techniques. Finally, we generate a cultural map by visualizing these cultural clusters on the map. Based on a large-scale user check-in dataset collected from Foursquare, we conduct both qualitative and quantitative

evaluation of the proposed approach. The results show that our approach can subtly capture cultural information from user behavioral data in LBSNs, and create representative cultural maps. Comparing our cultural maps with those created by traditional cultural mapping approaches based on survey data, we observe not only important cultural correlations between them, but also interesting differences caused by some unique cultural features extracted from user behavioral data.

In the future, since different parts of a city may include diverse cultures (e.g., China town and Wall street in New York), we plan to explore cultural maps with a different geographical granularity, such as different districts in a city. In addition, in order to augment user behavioral data in cultural mapping, we would like to capture and compare user behavior in different LBSNs, such as Twitter and Facebook, etc. Finally, as we continuously collect social media data, we plan to explore cultural diffusion by studying long-term user activity data in LBSNs.

The work in this chapter has not been previously published.

Chapter 8

Reflections and Outlook

Contents

Human dynamics, which focuses on understanding individual and collective human behavior, has been a core subject of study in various disciplines, such as psychology, sociology and physics. For example, in the long history of human development, psychologists, social scientists and urbanists have been theorizing with physical models to explain individual decision making process, information diffusion in society, human migration trends, urban commuting patterns, etc. Despite of extensive studies in understanding human dynamics, there has been a lack of large-scale datasets, which becomes a main limitation in studying human dynamics.

The rise of social media in the past decades has brought a revolution to large-scale empirical studies of human dynamics. The recent popularity of location-centric social media is expected to bring new impetus for exploring both individual and collective behavior. Specifically, millions of digital footprints that are left by a large number of users everyday provide an unprecedented opportunity to exploring large-scale human activity, and constitute a novel primary resource for the development of new applications of services.

In this dissertation, using large-scale user activity data from location centric social media, we have taken a step forward in studying human dynamics from both individual and collective perspectives, and exploring the potential applications of such knowledge.

8.1 Thesis Summary and Contributions

The contribution of this thesis involves the whole life-circle of the research process, including data collection, analysis and applications.

First, in order to obtain large-scale human activity datasets, in Chapter 3, we present a scalable data collection platform that we developed within the SOCIETIES project. The proposed platform is able to continuously collect global user activity data from different location centric social media in a streaming manner. Moreover, according to the specific characteristics of user activity data in LBSNs, we propose several noisy data filtering steps.

Second, aiming at exploring human dynamics from individual perspective, based on city-scale user activity data in LBSNs, we explore user preference on POIs and spatial temporal regularity of user activities. Specifically,

— In Chapter 4, aiming at studying user preference on POIs, we define two types of user preference, i.e., coarse-grained user preference (i.e., user-POI preference) and fine-grained user preference (i.e., user-POI-item preference), from heterogeneous user activity data in LBSNs (e.g., check-ins and user's comments). Afterwards, by investigating the characteristics of each types of user preference, we explore their applications in personalized location based services and propose a preference-aware POI recommendation and search framework. Specifically, by formulating the personalized recommendation and search tasks as user preference prediction problems, we propose two novel algorithms (i.e., LBSFM and MT-RTF algorithms) based on low-rank approximation techniques. The experimental evaluation shows that our framework can subtly capture user preference, and efficiently deliver personalized recommendation and search services.

— In Chapter 5, in order to explore the spatial temporal patterns of user activities in LBSNs, we propose STAP model for spatial temporal user activity preference modeling. For the spatial pattern, by discovering the spatial specificity property of user activity, we propose the concept of personal functional region to model and infer user spatial activity preference. For the temporal pattern, we propose to exploit the temporal correlation of user activities, and apply non-negative tensor factorization techniques to collaboratively infer user temporal activity preference. Finally, we put forward a context-aware fusion framework to combine the spatial and temporal models for activity preference inference tasks. The experimental evaluation proves that our model can effectively capture the spatial temporal regularity of user activities and outperform state-of-the-art solutions in the user activity preference inference task.

Third, aiming at exploring human dynamics from collective perspective, based on global-scale user activity data in LBSNs, we explore the collective activity pattern with

both country and city granularity, and its correlation with global cultures.

— In Chapter 6, we explore global-scale nation-wide collective activities in LBSNs. Specifically, we design and develop NationTelescope, a platform that monitors, compares, and visualize large-scale collective behavior in LBSNs. Via this platform, we are able to explore behavioral differences across countries, which often reflect cultural differences between countries. By implementing a prototype of NationTelescope platform, we evaluate its effectiveness and usability via two case studies and a System Usability Scale survey. The results show that the platform cannot only efficiently capture, compare and visualize nation-wide collective behavior, but also achieve good usability and user experience.

— In Chapter 7, in order to discover global cultures from collective behavior in LBSNs, we explore global-scale city-wide collective activities in LBSNs and their correlation with various cultural factors, such as geography, immigration and religion, etc. We propose a participatory cultural mapping approach to cluster cities into cultural clusters and plot a world cultural map with city granularity. Specifically, the proposed approach first eliminates non-local users in cities, whose activities are considered to be ineligible for characterizing local culture. Afterwards, by extracting three key cultural features from daily activity, mobility and linguistic perspectives respectively, we propose a cultural clustering method based on spectral clustering techniques to discover cultural clusters. The experimental results shows that our approach can efficient capture cultural features from collective activities in LBSNs, and generate representative cultural maps.

8.2 Directions of Future Research

In this dissertation, we investigate human dynamics and its applications using large-scale user activity data from location centric social media. Our study shows that such user activity data not only reflect individual preference and user daily activity patterns, but also contains cultural information with respect to collective behavior patterns. In the following, by pointing out several limitations of our work, we discuss some promising future research directions.

8.2.1 Data Fusion from Heterogeneous Sources

As users voluntarily report their activities in LBSNs, such data is not a full representation of user daily activity, and is thus a biased sample. While the primary goal of a user checking in at a POI is to share her real-time presence within her social circle, the user activity data in LBSNs can be regarded as a *social representation* of their daily activities.

Moreover, the user community of LBSNs is not a unified sample of the whole population. The user community of LBSNs is probably biased towards young people who prefer to use social media frequently. Therefore, we are aware that such biased data cannot fully reflect all aspects of human dynamics.

In order to augment user activity data from LBSNs, a promising research direction is to combine user digital traces from different social media and considering more data modality. For example, Flickr, which is a photo sharing social media, attracts many users to post their photos when travelling. Using geo-tagged photo sharing records in Flickr, Kurashima et al. [71] studied the travel route recommendation problem. In addition, photos further encompass rich semantic information about user activities. For example, by extracting and incorporating the semantic information of photos in LBSNs, Zhao et al. [161] studied the overlapping community detection profiling problem. We believe that the combination of heterogeneous data modality across different social media can reveal more "hidden facts" of human behavior, and thus better understand human dynamics.

In addition to social media, with the rapid advances of sensing technology, more data source in multiple domains become available (e.g., open data [8]), which can further augment user activity data. For example, in transportation domain, Metropolitan Transportation Authority (MTA) of the state of New York publishes the real-time traffic data of all subways in the city [1]; in public safety domain, Metropolitan Police Department of Washington D.C. publishes the historical crime incident data and its real-time feed [2]; in environment domain, New York city publishes the noise complaints data in the city [3]. By combining these data sources with LBSNs, more interesting aspects of human dynamics can be discovered. For example, Zheng et al. [163] combined user activity data in LBSNs with various data sources, such as such as noise compliant data, traffic data and road networks, to study the urban noise categorization problem.

8.2.2 Privacy Protection

The study of human dynamics is usually based on human activity data, which often concerns user privacy. For example, the user adoption of LBSNs is often hindered by growing user concern about privacy [39, 146]. Therefore, it is necessary to protect user privacy when studying human dynamics, particularly when enabling applications such as personalized services. In academia, researchers have started to study the privacy preserving personalization [98]. In most cases, privacy protection and personalization is contradictory. In other word, more information we know about a user, better personalization we can

1. http://web.mta.info/developers/
2. http://crimemap.dc.gov/
3. https://data.cityofnewyork.us/Social-Services/noise/xwca-wcf8

provide to her. Following this direction, Salamatian et al. investigated the trade-off between privacy protection and personalization performance [121]. As users continuously report their activities in LBSNs, the privacy protection becomes a more and more important research direction.

8.2.3 Big Human Activity Data

With the growing adoption of location centric social media, users have left a tremendous volume of data, and generate a considerable volume of activity data everyday. For example, Foursquare has attracted more than 45 million users globally and contained more than 5 billion check-ins by January 2014, with millions more every day. Therefore, it is necessary to explore new ways of accommodating these accumulated data from LBSNs. As a typical big data scenario, the four "V" (i.e., Volume, Velocity, Variety and Veracity) issues of such big human activity data need to be explored. In this dissertation, by studying human dynamics from the big human activity data, we partially tackle the variety and veracity issue. From data variety perspective, in Chapter 4, we explore user preference on POIs with heterogeneous data format (i.e., check-ins and tips); in Chapter 5, we study spatial and temporal pattern of user activity. From data veracity perspective, in Chapter 3, we filter out the noisy check-ins from a specific types of malicious users, i.e., "sudden-move" users; in Chapter 7, targeting to a specific application scenario, i.e., cultural mapping, we identify the local users who are representative to characterize the local culture.

In the future, we plan to explore more about the volume and velocity issues. For example, from data volume perspective, users, POIs and other data modality in LBSNs naturally construct a very large-scale hypergraph [161]. In order to mining valuable information from such a graph, novel efficient distributed graph mining algorithms may probably be needed. Moreover, from data velocity perspective, as user activity data continuously accumulates in a streaming manner, the streaming data processing techniques are required for efficient data processing.

8.3 Outlook

With the emergence and popularity of ubiquitous smart devices, location centric social media is becoming more and more popular and attracting an increasing number of users. These social media services pave the way for a broader trend: large-scale user activity data will be increasingly available. Such data massively implies various aspects of human behavior, which not only brings new opportunities to understand and explore human dynamics, but also comes with the challenges that concern the analysis and applications of such big social media data.

This dissertation has made a step forward in studying human dynamics from user activity data in location centric social media. As more user activity data is becoming available, our findings and results open the door to a vast range of future research directions. Besides the strictly quantitative aspects in this dissertation that may be volatile in light of future experimentations with novel data sources or methodologies, we hope that this work and its results can facilitate not only researchers in various academic disciplines, but also practitioners in the area when building new applications.

Appendix A

Appendix

A.1 Proof of Proposition 1

In one frequented region r of a user u, the number of user visited location categories $|C_{u,r}|$ is less than or equal to that of the existing categories $|C_l^d|$, i.e., $|C_{u,r}| \leq |C_l^d|$. The maximum entropy of a variable x is when x follows the uniform distribution. Thus, we have

$$H_{max}(C_{u,r}) = \log_2 |C_{u,r}| \leq H_{max}(|C_l^d|) = \log_2 |C_l^d| \tag{A.1}$$

The entropy of user u's actual visit distribution $H(\psi_{u,r})$ is less than or equal to that of uniform visit distribution $H_{max}(C_{u,r})$. Thus, we obtain

$$H(\psi_{u,r}) \leq H_{max}(C_{u,r}) \leq H_{max}(|C_l^d|) \tag{A.2}$$

Since entropy is non-negative, we have

$$0 \leq \frac{H(\psi_{u,r})}{H_{max}(|C_l^d|)} \leq 1 \tag{A.3}$$

Bibliography

[1] Center for human dynamics in the mobile age. `http://humandynamics.sdsu.edu/index.html`. Accessed: 2014-09-12.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[3] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási. Flavor network and the principles of food pairing. *Scientific reports*, 1, 2011.

[4] L. Ana and A. K. Jain. Robust data clustering. In *Proceedings of the IEEE International Conference Computer Vision and Pattern Recognition*, pages 128–133. IEEE, 2003.

[5] P. Anick. Using terminological feedback for web search refinement: a log-based study. In *Proceedings of the 26th ACM SIGIR International Conference on Research and Development in Informaion Retrieval*, pages 88–95. ACM, 2003.

[6] J. Antikainen. The concept of Functional Urban Area. *Informationen zur Raumentwicklung*, pages 447–456, 2005.

[7] Q. Ashraf and O. Galor. Cultural diversity, geographical isolation, and the origin of the wealth of nations. Technical report, National Bureau of Economic Research, 2011.

[8] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[9] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Language Resources and Evaluation Conference*, volume 10, pages 2200–2204, 2010.

[10] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[11] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 199–208. ACM, 2012.

[12] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[13] B. M. Bass. *Leadership, psychology, and organizational behavior.* Harper, 1960.

[14] S. Bauer, A. Noulas, D. O. Séaghdha, S. Clark, and C. Mascolo. Talking places: Modelling and analysing linguistic content in foursquare. In *Proceedings of the International Confernece on Social Computing*, pages 348–357. IEEE, 2012.

[15] P. Bedi, H. Kaur, and S. Marwaha. Trust based recommender system for semantic web. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 2677–2682, 2007.

[16] J. Bell. The world's muslims: Unity and diversity, 2012.

[17] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, pages 49–62. ACM, 2009.

[18] P. Berg, E. Appelbaum, T. Bailey, and A. L. Kalleberg. Contesting time: International comparisons of employee control of working time. *Industrial and Labor Relations Review*, pages 331–349, 2004.

[19] B. Berjani and T. Strufe. A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems*, page 4. ACM, 2011.

[20] M. H. Bond and K. Leung. *Cultural Mapping of Beliefs About the World and Their Application to a Social Psychology Involving Culture.* Taylor & Francis, 2009.

[21] M. R. Bouadjenek, H. Hacid, and M. Bouzeghoub. Sopra: A new social personalized ranking function for improving web search. In *Proceedings of the 36th ACM SIGIR International conference on Research and Development in Information Retrieval*, pages 861–864. ACM, 2013.

[22] J. Brooke. Sus-a quick and dirty usability scale. *Usability Evaluation in Industry*, 189:194, 1996.

[23] G. G. Brown. Missions and cultural diffusion. *American Journal of Sociology*, pages 214–219, 1944.

[24] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *Proceedings of the Workshop on World-Sensor-Web*, pages 117–134, 2006.

[25] Y. Cai and Q. Li. Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 969–978. ACM, 2010.

[26] X. Cao, G. Cong, and C. S. Jensen. Mining significant semantic locations from gps data. *Proceedings of the VLDB Endowment*, 3(1-2):1009–1020, 2010.

[27] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the 10th International Workshop on Multimedia Data Mining*, pages 4–14. ACM, 2010.

[28] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web*, pages 721–730. ACM, 2009.

[29] J. Chang and E. Sun. Location3: How users share and respond to location-based data on social networking sites. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 74–80, 2011.

[30] H.-C. Chen and A. L. Chen. A music recommendation system based on music data grouping and user interests. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, pages 231–238. ACM, 2001.

[31] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 12, pages 17–23, 2012.

[32] H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In *Proceedings of the 24th IEEE International Conference on Data Engineering*, pages 169–178. IEEE, 2008.

[33] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. volume 2011, pages 81–88, 2011.

[34] P.-A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the web. In *Proceedings of the 30th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 7–14. ACM, 2007.

[35] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1082–1090. ACM, 2011.

[36] D.-Y. Choi. Personalized local internet in the location-based mobile web search. *Decision Support Systems*, 43(1):31–45, 2007.

[37] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[38] M. Cole and J. S. Bruner. Cultural differences and inferences about psychological processes. *American Psychologist*, 26(10):867, 1971.

[39] P. Coppens, L. Claeys, C. Veeckman, and J. Pierson. Privacy in location-based social networks: Researching the interrelatedness of scripts and usage. In *Proceedings of the Symposium on Usable Privacy and Security*, 2014.

[40] C. Counihan and P. Van Esterik. *Food and culture: A reader*. Routledge, 2012.

[41] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, pages 58–65, 2012.

[42] L. De Lathauwer, B. De Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[43] T. M. T. Do and D. Gatica-Perez. Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 163–172. ACM, 2012.

[44] P. Du Gay and M. Pryke. *Cultural economy: cultural analysis and commercial life*. Sage, 2002.

[45] M. Eirinaki and M. Vazirgiannis. Web mining for web personalization. *ACM Transactions on Internet Technology*, 3(1):1–27, 2003.

[46] G. Evans and J. Foord. Cultural mapping and sustainable communities: planning for the arts revisited. *Cultural trends*, 17(2):65–96, 2008.

[47] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 114–121, 2012.

[48] A. Garzón, G. A. Cano, and G. Poussin. *Culture, trade and globalization: questions and answers*. Division of Creativity, Cultural Industries and Copyright, Sector for Culture, UNESCO, 2003.

[49] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[50] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[51] B. Guo, D. Zhang, and D. Yang. Read more from business cards: toward a smart social contact management system. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 384–387. IEEE Computer Society, 2011.

[52] B. Guo, D. Zhang, D. Yang, Z. Yu, and X. Zhou. Enhancing memory recall via an intelligent social contact management system. *IEEE Transactions on Human-Machine Systems*, 44(1):78–91, 2014.

[53] V. Gupta, P. J. Hanges, and P. Dorfman. Cultural clusters: Methodology and findings. *Journal of world business*, 37(1):11–15, 2002.

[54] F. A. Hansen, N. O. Bouvin, B. G. Christensen, K. Grønbæk, T. B. Pedersen, and J. Gagach. Integrating the web and the world: contextual trails on the move. In *Proceedings of the ACM Conference on Hypertext and Social Media*, pages 98–107. ACM, 2004.

[55] C. A. Heatwole. Culture: A geographical perspective, 2006.

[56] E. A. Hoebel. *Anthropology: The study of man.* McGraw-Hill New York, 1972.

[57] G. Hofstede. Cultural differences in teaching and learning. *International Journal of intercultural relations*, 10(3):301–320, 1986.

[58] R. Inglehart and C. Welzel. Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics*, 8(02):551–567, 2010.

[59] M. Iwata, T. Hara, K. Shimatani, T. Mashita, K. Kiyokawa, S. Nishio, and H. Takemura. A location-based content search system considering situations of mobile users. *Procedia Computer Science*, 5:426–433, 2011.

[60] M. Jamali and M. Ester. Trustwalker: a random walk model for combining trust-based and item-based recommendation. In *Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, pages 397–406. ACM, 2009.

[61] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the 4th ACM Conference on Recommender Systems*, pages 135–142. ACM, 2010.

[62] M. Jamali and M. Ester. A transitivity aware matrix factorization model for recommendation in social networks. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 2644–2649, 2011.

[63] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 793–801. ACM, 2013.

[64] A. Kawaguchi. Why is japanese working time is so long?: Wage-working time contract models. *Japanese economic review*, 47(3):251–270, 1996.

[65] H.-N. Kim, M. Rawashdeh, A. Alghamdi, and A. El Saddik. Folksonomy-based personalized search and ranking in social media services. *Information Systems*, 37(1):61–76, 2012.

[66] J. Kim and H. Park. Fast nonnegative tensor factorization with an active-set-like method. *High-Performance Scientific Computing*, pages 311–326, 2012.

[67] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[68] C. Kramsch. *Language and culture*. Oxford University Press, 1998.

[69] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

[70] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura. Geo topic model: joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pages 375–384. ACM, 2013.

[71] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 579–588. ACM, 2010.

[72] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[73] K. N. Laland, J. Odling-Smee, and S. Myles. How culture shaped the human genome: bringing genetics and the human sciences together. *Nature Reviews Genetics*, 11(2):137–148, 2010.

[74] N. D. Lane, D. Lymberopoulos, F. Zhao, and A. T. Campbell. Hapori: context-based local search for mobile phones using community behavioral modeling and similarity. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 109–118. ACM, 2010.

[75] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *In Proceedings of Mobile Data Challenge by Nokia Workshop*, 2012.

[76] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[77] K. Leetaru, S. Wang, G. Cao, A. Padmanabhan, and E. Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.

[78] P. Levitt. Social remittances: migration driven local-level forms of cultural diffusion. *International migration review*, pages 926–948, 1998.

[79] J. R. Lewis and J. Sauro. The factor structure of the system usability scale. In *Human Centered Design*, pages 94–103. Springer, 2009.

[80] D. Lian and X. Xie. Collaborative activity recognition via check-in history. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 45–48. ACM, 2011.

[81] Y. Liang, J. Caverlee, Z. Cheng, and K. Y. Kamath. How big is the crowd?: event and location based population modeling in social media. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 99–108. ACM, 2013.

[82] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55, 1932.

[83] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[84] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2409–2418. ACM, 2011.

[85] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*, pages 63–70. Association for Computational Linguistics, 2002.

[86] H. Ma, I. King, and M. R. Lyu. Learning to recommend with social trust ensemble. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 203–210. ACM, 2009.

[87] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pages 287–296. ACM, 2011.

[88] M. Madden, S. Fox, A. Smith, and J. Vitak. *Digital Footprints: Online identity management and search in the age of transparency.* Pew Internet & American Life Project Washington, DC, 2007.

[89] P. A. Madden, A. C. Heath, N. E. Rosenthal, and N. G. Martin. Seasonal changes in mood and behavior: the role of genetic factors. *Archives of General Psychiatry*, 53(1):47–55, 1996.

[90] T. Maekawa, Y. Yanagisawa, Y. Sakurai, Y. Kishino, K. Kamei, and T. Okadome. Context-aware web search in ubiquitous sensor environments. *ACM Transactions on Internet Technology*, 11(3):12, 2012.

[91] P. Massa and P. Avesani. Trust-aware recommender systems. In *Proceedings of the ACM Conference on Recommender Systems*, pages 17–24. ACM, 2007.

[92] W. C. McGrew. Culture in nonhuman primates? *Annual Review of Anthropology*, 27(1):301–328, 1998.

[93] M. Mehaffy, S. Porta, Y. Rofè, and N. Salingaros. Urban nuclei and the geometry of streets: The 'emergent neighborhoods' model. *Urban Design International*, 15(1):22–46, 2010.

[94] A. Mnih and R. Salakhutdinov. Probabilistic matrix factorization. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1257–1264, 2007.

[95] R. T. Moran, P. R. Harris, and S. Moran. *Managing cultural differences*. Routledge, 2007.

[96] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering analysis and an algorithm. *Proc. NIPS*, 14:849–856, 2001.

[97] W. Ng, L. Deng, and D. L. Lee. Mining user preference using spy voting for search engine personalization. *ACM Transactions on Internet Technology*, 7(4):19, 2007.

[98] V. Nikolaenko, S. Ioannidis, U. Weinsberg, M. Joye, N. Taft, and D. Boneh. Privacy-preserving matrix factorization. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security*, pages 801–812. ACM, 2013.

[99] A. Noulas, C. Mascolo, and E. Frias-Martinez. Exploiting foursquare and cellular data to infer user activity in urban environments. In *Proceedings of the 14th IEEE International Conference on Mobile Data Management*, volume 1, pages 167–176. IEEE, 2013.

[100] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[101] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pages 1038–1043, 2012.

[102] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Proceedings of the International Confernece on Social Computing*, pages 144–153. IEEE, 2012.

[103] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. volume 11, pages 570–573, 2011.

[104] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. *The Social Mobile Web*, 11, 2011.

[105] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, pages 167–174. ACM, 2005.

[106] A. D. of Communications and the Arts. *Mapping culture : a guide for cultural and economic development in communities*. Canberra : A.G.P.S, 1995.

[107] E. Ostrom. Collective action and the evolution of social norms. *The Journal of Economic Perspectives*, pages 137–158, 2000.

[108] J. Park, V. Barash, C. Fink, and M. Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 466–475, 2013.

[109] A. Pentland, T. Choudhury, N. Eagle, and P. Singh. Human dynamics: computation for organizations. *Pattern Recognition Letters*, 26(4):503–511, 2005.

[110] C. Perreault and P. J. Brantingham. Mobility-driven cultural transmission along the forager–collector continuum. *Journal of Anthropological Archaeology*, 30(1):62–68, 2011.

[111] F. Pianese, X. An, F. Kawsar, and H. Ishizuka. Discovering and predicting user routines by differential analysis of social network traces. In *Proceedings of the 14th IEEE International Symposium and Workshops on a World of Wireless, Mobile and Multimedia Networks*, pages 1–9. IEEE, 2013.

[112] P. Poole. Cultural mapping and indigenous peoples. *A report for UNESCO*, 2003.

[113] D. Preoţiuc-Pietro, J. Cranshaw, and T. Yano. Exploring venue-based city-to-city similarity measures. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 16. ACM, 2013.

[114] E. G. Ravenstein. The laws of migration. *Journal of the Statistical Society of London*, pages 167–235, 1885.

[115] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 727–736. ACM, 2009.

[116] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press, 2009.

[117] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, pages 81–90. ACM, 2010.

[118] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the ACM conference on Computer supported cooperative work*, pages 175–186. ACM, 1994.

[119] R. Robertson. *Globalization: Social theory and global culture*, volume 16. Sage, 1992.

[120] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM, 2010.

[121] B. K. Salman Salamatian, Nadia Fawaz and N. Taft. Sspm: Sparse privacy preserving mappings. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 712–721, 2014.

[122] J. Sam and S.-D. Olivia. Uk v france: The sunday shopping difference. *BBC News Magazine*, 2013.

[123] J. Sang, C. Xu, and D. Lu. Learn to personalized image search from the photo sharing websites. *IEEE Transactions on Multimedia*, 14(4):963–974, 2012.

[124] E. Sapir. Language as a form of human behavior. *The English Journal*, 16(6):421–433, 1927.

[125] S. H. Schwartz. Mapping and interpreting cultural differences around the world. *International studies in sociology and social anthropology*, pages 43–73, 2004.

[126] S. Seagal and D. Horne. *Human dynamics: A new framework for understanding people and realizing the potential in our organizations.* Pegasus Communications Waltham, MA, 1997.

[127] Y. Shi, A. Karatzoglou, L. Baltrunas, M. Larson, A. Hanjalic, and N. Oliver. Tfmap: optimizing map for top-n context-aware recommendation. In *Proceedings of the 35th ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 155–164. ACM, 2012.

[128] N. Shuyo. Language detection library for java, 2010.

[129] T. H. Silva, P. Vaz De Melo, J. M. Almeida, and A. A. Loureiro. Large-scale study of city dynamics and urban social behavior using participatory sensing. *IEEE Wireless Communications*, 21(1):42–51, 2014.

[130] B. F. Skinner. *Science and human behavior.* Simon and Schuster, 1953.

[131] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[132] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th International Conference on World Wide Web*, pages 675–684. ACM, 2004.

[133] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. Cubesvd: a novel approach to personalized web search. In *Proceedings of the 14th International Conference on World Wide Web*, pages 382–390. ACM, 2005.

[134] S. Tarrow and Tollefson. *Power in movement: Social movements, collective action and politics*. Cambridge Univ Press, 1994.

[135] W. Taylor. *A study of archeology*. Arcturus books. Southern Illinois University Press, 1967.

[136] H. C. Triandis. The self and social behavior in differing cultural contexts. *Psychological review*, 96(3):506, 1989.

[137] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[138] R. H. Turner and L. M. Killian. *Collective behavior*. Prentice-Hall, 1957.

[139] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[140] K. Waga, A. Tabarcea, and P. Fränti. Context aware recommendation of location-based data. In *Proceedings of the 15th International Conference on System Theory, Control, and Computing*, pages 1–6, 2011.

[141] Z. Wang, D. Zhang, D. Yang, Z. Yu, and X. Zhou. Detecting overlapping communities in location-based social networks. In *Proceedings of the 2012 International Conference on Social Informatics*, pages 110–123. Springer, 2012.

[142] Z. Wang, D. Zhang, D. Yang, Z. Yu, X. Zhou, and Z. Yu. Investigating city characteristics based on community profiling in lbsns. In *Proceedings of the 2012 Second International Conference on Cloud and Green Computing*, pages 578–585. IEEE, 2012.

[143] Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, and Z. Yu. Discovering and profiling overlapping communities in location-based social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(4):499–509, April 2014.

[144] Z. Wang, X. Zhou, D. Zhang, D. Yang, and Z. Yu. Cross-domain community detection in heterogeneous social networks. *Personal and ubiquitous computing*, 18(2):369–383, 2014.

[145] J. Weng and B.-S. Lee. Event detection in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 401–408, 2011.

[146] J. Xie, B. P. Knijnenburg, and H. Jin. Location sharing privacy preference: Analysis and personalized recommendation. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, pages 189–198. ACM, 2014.

[147] S. Xu, S. Bao, B. Fei, Z. Su, and Y. Yu. Exploring folksonomy for personalized search. In *Proceedings of the 31st ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 155–162. ACM, 2008.

[148] D. Yang, D. Zhang, K. Frank, P. Robertson, E. Jennings, M. Roddy, and M. Lichtenstern. Providing real-time assistance in disaster relief by leveraging crowdsourcing power. *Personal and Ubiquitous Computing*, 18(8):2025–2034, 2014.

[149] D. Yang, D. Zhang, Z. Yu, and Z. Wang. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 119–128. ACM, 2013.

[150] D. Yang, D. Zhang, Z. Yu, and Z. Yu. Fine-grained preference-aware location search leveraging crowdsourced digital footprints from lbsns. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 479–488. ACM, 2013.

[151] D. Yang, D. Zhang, Z. Yu, Z. Yu, and D. Zeghlache. Sesame: Mining user digital footprints for fine-grained preference-aware social media search. volume 14, pages 28:1–28:24. ACM, 2014.

[152] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015.

[153] J. Ye, Z. Zhu, and H. Cheng. What's your next move: User activity prediction in location-based social networks. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2013.

[154] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 458–461. ACM, 2010.

[155] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 325–334. ACM, 2011.

[156] Z. Yu, D. Zhang, and D. Yang. Where is the largest market: Ranking areas by popularity from location based social networks. In *Proceedings of the 10th IEEE International Conference on Ubiquitous Intelligence and Computing*, pages 157–162. IEEE, 2013.

[157] Z. Yu, D. Zhang, D. Yang, and G. Chen. Selecting the best solvers: toward community based crowdsourcing for disaster management. In *Proceedings of the 2012 IEEE Asia-Pacific Services Computing Conference*, pages 271–277. IEEE, 2012.

[158] Z. Yu, D. Zhang, Z. Yu, and D. Yang. Participant selection for offline event marketing leveraging location-based social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2015. to be published.

[159] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 186–194. ACM, 2012.

[160] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie. We know how you live: exploring the spectrum of urban lifestyles. In *Proceedings of the first ACM conference on Online social networks*, pages 3–14. ACM, 2013.

[161] Y.-L. Zhao, Q. Chen, S. Yan, T.-S. Chua, and D. Zhang. Detecting profilable and overlapping communities with user-generated multimedia contents in lbsns. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 10(1):3, 2013.

[162] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Ciscovery and Data Mining*, pages 1436–1444. ACM, 2013.

[163] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, and E. CHANG. Diagnosing new york city's noises with ubiquitous data. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 715–725. ACM, 2014.

[164] D. Zhou, S. Lawless, and V. Wade. Web search personalization using social data. In *Theory and Practice of Digital Libraries*, pages 298–310. Springer, 2012.

[165] Y.-X. Zhu, J. Huang, Z.-K. Zhang, Q.-M. Zhang, T. Zhou, and Y.-Y. Ahn. Geography and similarity of regional cuisines in china. *PloS one*, 8(11):e79161, 2013.